

STATISTICS FOR BUSINESS DECISIONS

STATISTICS FOR BUSINESS DECISIONS

By

ERNEST KURNOW

GERALD J. GLASSER

and

FREDERICK R. OTTMAN

ALL OF NEW YORK UNIVERSITY

SCHOOL OF COMMERCE, ACCOUNTS, AND FINANCE AND

GRADUATE SCHOOL OF BUSINESS ADMINISTRATION



1959

RICHARD D. IRWIN, INC.

HOMEWOOD, ILLINOIS

311.2
K96A

© 1959 BY RICHARD D. IRWIN, INC.

ALL RIGHTS RESERVED. THIS BOOK OR ANY PART THEREOF MAY NOT
BE REPRODUCED WITHOUT THE WRITTEN PERMISSION OF THE PUBLISHER

First Printing, June, 1959
Second Printing, September, 1959
Third Printing, August, 1960
Fourth Printing, August, 1961
Fifth Printing, February, 1962

Library of Congress Catalogue Card No. 59-11219

PRINTED IN THE UNITED STATES OF AMERICA

Preface

This book deals with the field of statistics and its use in the making of decisions. It represents an attempt to present a college textbook on elementary statistics from what may be called a behavioralistic point of view.¹ In this respect and in several others as well we feel that this book differs from any other introductory text in the field.

The behavioralistic outlook regards statistics as a medium for deciding upon an alternative course of action. Thus, the emphasis in this book is on how to formulate a decision problem statistically so that data, when collected, will serve as a basis for deciding upon a rational course of action. The collection of data is, of course, never an end in itself—at least it never should be an end in itself. In our opinion, this is the most important idea that an elementary course in statistics should convey to a student. Hence, we have attempted to make the relationship of data to decisions a central theme in the organization and presentation of material. Stress is laid not on the mere details of gathering statistical information, but on how such information is to be used once it is collected.

In a word, then, this book emphasizes *planning*—planning statistical studies so that when conducted they provide a basis for decision. With this principle of planning as a foundation stone, other important aims of our text include the following:

- a) Developing in the student a skeptical attitude toward numerical data.
- b) Impressing upon the student the necessity for looking behind figures to the procedure used to derive them.
- c) Introducing a way of thinking, not only in terms of numbers, but in terms of uncertainty and risk.
- d) Indicating to the student that statistics is of ever increasing importance in the field of business by presenting some of the many applications of business statistics.

In so far as possible, material is presented to stress ideas and concepts and principles, rather than computations. Furthermore,

¹ The term "behavioralistic" is borrowed, with essentially the same meaning, from Leonard J. Savage, *The Foundations of Statistics* (New York: John Wiley & Sons, Inc., 1954).

the approach is essentially nonmathematical. Techniques are treated only when they further statistical reasoning, and not as ends in themselves. The goal in this connection is to first bring the student to the point where he knows what statistics can do. Only where it does not detract from this objective is he taught the fine details of "how to do it." As an aid in achieving these goals we include more than 250 examples of decision problems which can be solved through the use of statistics. These examples are an integral part of the text, and the student must be encouraged not only to read them, but to analyze them thoroughly. Our belief is that statistical concepts are best clarified by specific illustrations and that these examples motivate as well as instruct.

With one or two minor exceptions only large sample statistical theory is treated throughout this text. For those who wish to illustrate the thought that other theoretical distributions are useful, Appendix E very briefly presents t and F .

The book may be viewed as consisting of two parts. The first part is represented by Chapters 1 to 10, which may be regarded as forming an introduction to statistics. These ten chapters are arranged to simulate the steps needed to plan a study for a decision problem which can be treated statistically. Thus, Chapter 1 treats material on identifying the problem, while Chapters 2 and 3 follow with ideas on how to define the problem in statistical terms. Chapter 4 focuses on alternative methods of collecting information (secondary sources, census, sample, etc.). Chapters 5 through 8 then center attention on how to formulate plans for sample studies. Chapter 9 warns against the imperfections that one may find in data. Finally, Chapter 10 deals with how one collects data after all of the plans for a study have been completed. Thus, we have postponed the collection, tabulation, and presentation of data until analysis (a priori analysis) of data has been treated fully. At that time the student should appreciate why data are being collected.

Chapters 11 through 14 form the second part of the book. Each chapter covers a special topic and each is an individual unit in itself. Alternative teaching plans for covering the material in the book are suggested in the teacher's manual.

Each chapter ends with a summary, a series of problems and questions, and a bibliography. Each summary is simply designed to provide a way for the student to verify whether he has absorbed the main ideas in the chapter. The problems are also designed to provide

a basis for verifying whether meaning and understanding have been achieved. Additional problems and questions that require more computation will be found in the workbook designed to accompany the text. The bibliography includes only references geared to the expected level of ability of most students, since it is very discouraging to a student to find references that he cannot cope with.

The three of us, and nine of our colleagues as well, have tried and tested drafts of *Statistics for Business Decisions* for six semesters at both the undergraduate and graduate levels. Approximately 2,000 students have used the text during that period. We have found their response gratifying. Furthermore, we have been able to make many adjustments and revisions in the original drafts on the basis of their comments. The rise in the number of students desiring to pursue the further study of statistics has provided additional psychic income.

In addition to our colleagues and students many other persons deserve our deepest and sincerest appreciation. Among these is Dean Thomas L. Norton who was instrumental in the introduction of a basic course in statistics required of all students at the School of Commerce, Accounts, and Finance, and who encouraged the writing of this textbook. His continuing interest hastened completion of the task.

To our good friend and colleague, Professor W. Edwards Deming, we owe many hours of time for reading and analyzing the manuscript. His suggestions were invariably valuable, and we only wish that we were able to incorporate them more adequately. Only those who know Dr. Deming and the heavy work-load that he carries can fully appreciate our obligation to him.

Dean Erwin A. Gaumnitz of the School of Commerce of the University of Wisconsin also read the manuscript, and his many suggestions concerning content, arrangement, and method of presentation have helped us to make the text more understandable. Professor Julius H. Spalding, our colleague for many years, reviewed the earlier chapters for clarity of presentation and content.

Mrs. Kay Degen typed the manuscript with amazing accuracy, and cheerfully and efficiently performed many other arduous tasks in connection with the development of this text. Miss Eve L. Crosby assisted us in reading the galley proofs and in typing the index.

Also of assistance were Ruth, Susan, and Alice Kurnow, and Mary Jane Ottman. Laura Glasser provided the basis for Example 1.16. Lastly, but not leastly, we must thank our wives, Joyce Kurnow,

Anne Glasser, and Dorothy Ottman for two years of patience and forbearance, as well as hours of assistance in the preparation of the book.

We are indebted to Professor Sir Ronald A. Fisher, Cambridge, and to Messrs. Oliver and Boyd, Ltd., Edinburgh, for permission to reprint Table IV (Table E.1. in this book) from their book *Statistical Methods for Research Workers*; to Professor Fisher and to Dr. Frank Yates, Rothamsted, and to Messrs. Oliver and Boyd, Ltd., for permission to reproduce a portion of Table V (the .001 values of Table E.2. of this book) from *Statistical Tables for Biological, Agricultural and Medical Research*. We are also indebted to Professor Egon Pearson and to *Biometrika* for permission to reproduce from Vol. XXXIII, the .05 and .01 values of F shown in Table E.2. of this book.

ERNEST KURNOW
GERALD J. GLASSER
FREDERICK R. OTTMAN

NEW YORK, N.Y.
May, 1959

Table of Contents

CHAPTER	PAGE
1. BUSINESS PROBLEMS, DECISIONS, AND STATISTICS	1
The Plague and Statistics, 1. Statistics and Statistics, 2. Decisions and Courses of Action, 3. States of the World, 6. Uncertainty, Observations, and Sampling, 10. Testing and Estimation Problems, 12. The Principle of Planning, 14. The Principle of Teamwork, 16. You Should Now Know That, 17. You Should Now Be Able to Solve, 18. You Will Also Find That, 19.	
2. STATISTICAL POPULATIONS	20
Defining the Problem Statistically, 20. Elementary Units and Their Characteristics, 22. The Frame, 25. Enumerative and Analytical Studies, 28. Multiobservational Problems, 31. Variates and Attributes, 32. Frequency Distributions, 36. Population Models, 41. You Should Now Know That, 46. You Should Now Be Able to Solve, 46. You Will Also Find That, 48.	
3. DECISION-PARAMETERS	50
Introduction, 50. The Proportion, 53. The Aggregate, 58. The Arithmetic Mean, 62. The Median and the Mode, 67. Variability in Populations, 71. The Standard Deviation and the Normal Curve, 77. Selecting the Decision-Parameter, 83. You Should Now Know That, 84. You Should Now Be Able to Solve, 85. You Will Also Find That, 87.	
4. SAMPLE SELECTION	88
Introduction, 88. Primary and Secondary Data, 90. Reasons for Sampling, 96. All Possible Samples, 100. Probability, Judgment, and Convenience Sampling, 102. Simple Random Sampling, 107. Random Number Tables, 111. Other Types of Probability Samples, 116. You Should Now Know That, 119. You Should Now Be Able to Solve, 120. You Will Also Find That, 121.	
5. PROBABILITY THEORY	123
Statistics and Statistics and Statistics, 123. Mathematical Probability, 127. Properties of a Probability Measure, 132. Repeated Trials—Conditional Probability, 134. Independent Repeated Trials, 141. The Binomial Formula, 144. Expected Values, 148. You Should Now Know That, 150. You Should Now Be Able to Solve, 151. You Will Also Find That, 153.	
6. SAMPLING DISTRIBUTIONS	154
Sampling Distributions Defined, 154. The Sampling Distribution of p , 156. Characteristics of the Sampling Distribution of p , 160. The Normal Curve and Sampling Risks, 165. Finite Twofold Populations, 170. The Sampling Distribution of the Mean, 175. Characteristics of the Sampling Distribution of $\bar{x}$, 184. Are There Sampling Distributions of Decision-Parameters? 189. You Should Now Know That, 190. You Should Now Be Able to Solve, 191. You Will Also Find That, 193.	
7. RISK AND UNCERTAINTY IN ESTIMATION PROBLEMS	194
Estimation Decision Problems, 194. Estimators, 198. Choice of a	

CHAPTER	PAGE
Method of Sampling, 207. Determining Sample Size, 210. The Precision Attained, 217. Interval Estimation, 220. Estimation of Distributions, 225. You Should Now Know That, 228. You Should Now Be Able to Solve, 228. You Will Also Find That, 230.	
8. RISK AND UNCERTAINTY IN TESTING PROBLEMS	232
Risks in Testing Problems, 232. Formulating the Decision Rule—the Proportion as a Decision-Parameter, 236. Formulating a Decision Rule—the Mean as a Decision-Parameter, 241. Operating Characteristic Curves, 248. Acceptance Sampling, 257. Other Acceptance Sampling Plans, 263. Management Objectives and Decision Rules, 268. You Should Now Know That, 271. You Should Now Be Able to Solve, 272. You Will Also Find That, 274.	
9. ERROR AND BIAS	275
Introduction: Sampling Errors versus Procedural Bias, 275. Improper Formulation of the Decision Problem, 277. Biased Selection Procedures, 279. Bias of Nonresponse, 282. Biased Observations, 286. Biases in Data Flow, 291. You Should Now Know That, 292. You Should Now Be Able to Solve, 293. You Will Also Find That, 294.	
10. DATA FLOW	296
Data Flow Defined, 296. Methods of Data Collection, 297. Organization of the Collection Procedure, 302. Processing the Data, 305. Tabulation of Data, 309. Computation of Statistics: The Mean, 313. Computation of Statistics: The Standard Deviation, 316. Machine Tabulation and Calculation, 319. The Final Report, 326. You Should Now Know That, 328. You Should Now Be Able to Solve, 329. You Will Also Find That, 331.	
11. CONTINUING DECISION PROBLEMS AND THE CONTROL CHART .	332
Introduction, 332. Process Variability, 334. The Control Chart for the Proportion Defective, 337. Control Charts for the Mean and Range, 344. The Control Chart for Number of Defects, 351. The Risks of Incorrect Action, 354. The Samples, Their Size, and Frequency of Selection, 357. Statistical Control and a Satisfactory Product, 360. You Should Now Know That, 363. You Should Now Be Able to Solve, 364. You Will Also Find That, 368.	
12. COMPARATIVE EXPERIMENTS AND THEIR STATISTICAL DESIGN .	369
The Nature of Experimentation, 369. Decisions Involving the Comparison of Two Population Means, 372. Decisions Involving the Comparison of Two Population Proportions, 379. Decisions Involving the Comparison of Two Populations—Design, 382. Decisions Involving the Comparison of Several Populations, 386. Decisions Involving the Comparison of Several Populations—Design, 391. You Should Now Know That, 396. You Should Now Be Able to Solve, 397. You Will Also Find That, 399.	
13. DECISIONS BY ASSOCIATION	401
Introduction, 401. Two Aspects of Association, 403. Association and Causation, 406. A Population Model for Regression Problems, 408. The Estimated Regression Line, 413. Sampling Error of Estimated Regression Line, 419. Judgment Applications of Regression, 428. The Population Coefficient of Correlation, 432. Sampling Theory for Correlation, 436. Rank Correlation, 438. You Should Now Know That, 441. You Should Now Be Able to Solve, 442. You Will Also Find That, 445.	

TABLE OF CONTENTS

xi

CHAPTER	PAGE
14. FORECASTING WITH TIME SERIES	446
Some Remarks on Forecasting, 446. Time Series and Forecasting, 447. Graphic Presentation of Time Series, 449. Nature of Economic Time Series, 453. The Analysis of the Trend Element, 457. The Analysis of Seasonal Movements, 470. Cyclical Fluctuations, 478. Trend, Seasonal, and Cyclical Elements—a Synthesis, 480. You Should Now Know That, 482. You Should Now Be Able to Solve, 483. You Will Also Find That, 485.	

APPENDICES

APPENDIX	
A. TABLE OF SQUARES, SQUARE ROOTS, AND RECIPROCAL, $N = 1$ TO 1,000	488
B. AREAS UNDER THE NORMAL CURVE AND NORMAL DEVIATES . .	500
C. TABLE OF 5,000 RANDOM DIGITS	503
D. TABLE OF COMMON LOGARITHMS, 100-999	506
E. TABLES OF THE t -DISTRIBUTION AND THE F -DISTRIBUTION . .	509
INDEX	515

CHAPTER • 1

Business Problems, Decisions, and Statistics

1.1 The Plague and Statistics

Throughout this book we shall be studying the theory and the methods of making rational decisions based upon numerical data. In particular, our concern will be with many of the problems arising in the modern business world.

As a simple starting point we can illustrate this general idea with an occurrence of great importance in the history of the subject we shall be studying—an occurrence which is nearing its 300th anniversary. It may be considered as an early illustration of decision making based upon numerical data.

Example 1.1 During the seventeenth century there were several outbreaks of the plague in the city of London, and apparently a large proportion of the city's inhabitants died during these unfortunate epidemics. Citizens followed announcements of births and, moreover, of deaths with considerable interest. At that time, in fact, summary information on the number of christenings and burials and records on the causes of death began to be published regularly.

Although rough and certainly crude, this numerical information was of some interest, and quite in demand by some of the public and by businessmen. For as one writer has interpreted history: "the well-to-do watched these data to note the threat of infection so as to leave London in good time, *while the merchants used them to draw business forecasts for the coming seasons.*"¹

¹Corrado Gini, "The First Steps of Statistics," *Educational Research Forum Proceedings* (International Business Machines Corporation, 1947), pp. 37-45. (Italics added.)

Perhaps history has been a little too conveniently interpreted here. Even so, make note of the idea of basing decisions in everyday affairs and in business upon numerical information. This will be the topic of central importance in our studies.

Despite this rather undistinguished beginning and the unfortunate circumstance of originally being associated with the plague, the idea of decision based upon numerical information is the basis of an important field of study today. This field is known as *statistics*. However, the word "statistics" represents an unfortunate name since it is quite ambiguous. For that reason we had better pass from the 1600's to the twentieth century and attempt to define it more precisely.

1.2 Statistics and Statistics

Although we have used the term "statistics" in the preceding section, we probably are using it with a slightly different meaning from that to which the student is accustomed. Let's try to clear up this possible difficulty. Think back to the first time that you heard or used that word. What did it refer to? Probably it was a very sophisticated way of referring to numbers, or data, or numerical information. This use of the word is perfectly all right, but it is only one of two meanings that we shall use.

In this first case, the word "statistics" refers to data such as the number of births and deaths, or the prices of scrap steel, or the wages of employees. As we shall see, statistical data of this type will be an essential ingredient in our studies, for data represent the information upon which statistical decisions are based. We should, in fact, regard numerical information as our raw material.

However, the subject that embraces the theory and the methods for making decisions from the data is also known as *statistics*. Therefore, in this second use of the word we are referring to a field of knowledge, and not merely to the data. Formalizing our present remarks and adding them to previous comments we can arrive at a more complete introductory definition of our subject—this field of statistics.

Let us say, for the present, that *statistics involves methods and theory as they are applied to numerical data or observations with the objective of making rational decisions in the face of uncertainty.*

This definition has implications which we will consider briefly for the moment and which we will review throughout the book. Firstly, the word "methods" refers to the statistical tools which are used to

make decisions on the basis of numerical information. In the course of our studies you will meet a number of these tools. Secondly, the word "theory" refers to the foundations of statistical methods. Put simply, the function of this theory is to make sure that proper tools are being used. Thirdly, the notion of making *decisions in the face of uncertainty* refers to the primary use of the methods and theory that we shall be studying.

Example 1.2 A pharmaceutical company wishes to decide on whether to market drug A, drug B, or neither drug. They plan to collect data on the effectiveness of the two drugs by an experiment on guinea pigs. That is, they plan to collect statistical information upon which to base their decision. They plan to collect statistics (in the first sense of the word).

How should these data be used to make the decision? How much data should they collect? What is the most effective way of running the experiment? What action should be taken on the basis of the data collected? What are the chances of their making the right marketing decision? These are all questions the field of statistics (the second use of the word) can help to answer.

Methods are the formulas or other techniques which will specifically answer these questions.

Theory is the basis for these methods—the guarantee, so to speak, that the techniques used are the correct ones.

The problem will have to be solved on the basis of incomplete information since there are limits to the number of guinea pigs that can be used and hence limits on the amount of data that can be collected. Incomplete information, in turn, means that there will be some *uncertainty* in the decision.

The problem falls within the realm of our subject.

In short, remember that we have two uses for the word "statistics." The first use simply indicates numerical data, while the second refers to a field of study. Grammatically, you should remember that in the first sense, the word is a plural noun (statistics are data), while in the second sense, the word is a singular noun (statistics is a subject). Although we shall make use of the plural (for raw materials), it is the singular with which this book will primarily deal.

1.3 Decisions and Courses of Action

In this book, we will, for the most part, restrict our discussion of the applications of statistics to business decision problems. This is only because our interests lie in this direction. This same subject, statistics, is also of considerable importance in many other fields, and similar methods and theory are also applicable there. Some of

these other fields are biology, medicine, psychology, agriculture, archaeology, and sociology. However, since most of our attention will be directed to the use of statistics in business and since statistics is concerned with decision, we now turn to survey, in the next few sections, some of the basic characteristics of the decision-making process in the business world. Our main interest and particular emphasis in doing so will be to see what role statistics plays in this process.

Suppose that we begin with a somewhat obvious but important point. Everyone makes some of his own personal decisions every day. What to wear? Where to eat? When to do this? How to do that? In business this same situation exists. Management must make certain decisions on business problems day after day. Some problems are minor; some are larger scale. Some are repetitive; some are relatively unique. Some are easily solved by snap decisions; some are more complicated.

Despite this diversity, all decision problems are characterized by two factors. First of all, an important element that must exist in any decision problem is the *principle of action*. By action, we mean that a decision problem is a problem in deciding on a course of action and not a matter of satisfying one's curiosity. A problem situation never calls for a decision unless action can be taken.

In this connection, a life insurance company recently gave out some good advice to its policyholders:

Don't let the unsolved problems of the day pursue you all night. As problems arise during the day, solve them. Make your decisions, then forget about them. Don't keep wrestling with problems that really aren't yours to solve. Most problems can be divided into two groups: those you can do something about and those over which you have no control . . . (The Connecticut Mutual Life Insurance Company.)

This advice is equally well suited for management in dealing with its business problems. Furthermore, this advice must go hand in hand with the use of statistics.

Statistics is concerned with many of those problems that you can do something about. If there is no control over the problem situation and no action can be taken, nothing can be done about the problem anyway. Obviously, unless action can be taken on a problem, there is no need to use statistics to help arrive at a solution for it. As you will see, if you haven't already, we would only be wasting time, effort, and money.

Statistics is useful in a large class of problems upon which man-

agement must make decisions, but statistics cannot be used to solve every problem. The point to remember here is that in decision making the tools of the statistician must be used as a basis for action if they are to be used profitably. *Statistics provides the basis for action.* If a businessman can't act on a problem, then statistics can't help him.

The second factor which characterizes every decision problem may be summarized by the term "alternatives." This refers to the fact that a decision problem exists only when more than one course of action is available. Specifying the existence of alternatives is simply another way of saying that there is a managerial problem. If management has but one course of action open to it, no problem actually exists and no decision is necessary. Management must take that course of action.

Thus, the term "decision" implies that management has various alternative courses of action open to it, and from these it selects one alternative.

When applicable, statistics provides rational means of selecting this one alternative and rejecting all others. This procedure gives rise to the terminology—*statistical decision theory*.

Suppose that we now turn to a few simple illustrations of business decision problems. For the present they should at least serve to stress how the concepts of action and alternatives appear in any decision-making situation. Ultimately we shall see how statistics can be used to help in solving these problems.

Example 1.3 A company's division of production research develops what it thinks is a better and less costly gadget than those presently available. Several problems may immediately arise; each calls for some managerial decision. Should the gadget be marketed? If so, what are the best markets? What is the best means of distribution? Of promotion? Of advertising? Each problem obviously has various alternatives associated with it. Incidentally, the company's market research division, using appropriate statistical methods, can provide reliable answers to these questions if management wishes to use statistics as a basis for decision and action.

Example 1.4 The section head of a company's bookkeeping department naturally wishes to have the error rate of his office's posting operation as low as possible, even though he realizes some mistakes are bound to occur occasionally. But he also wishes to keep the costs of doing this job at a minimum. Various alternatives are open to the section head, for he may choose the option of not checking any work done by the clerks, or he may choose to check all work done by

some clerks, or to check all work done by all clerks. He might, in fact, even decide to double-check or to triple-check all work.

It must be assumed, of course, that he could carry out each of these possibilities. If he is short of help, he might not be able to check any work—he might have only one course of action—not checking—open to him. In such a case, no decision problem exists.

Example 1.5 A manufacturer of oilcloth is faced with the problem of disposing of each roll as it is produced. The quality of a roll is measured by the average number of blemishes per square yard. The manufacturer considers three alternative courses of action: (1) to sell a roll as “first quality,” (2) to sell it as “second quality,” or (3) to scrap the roll. Thus the manufacturer's decision and his course of action may vary from roll to roll, but in each case his decision will be based on the quality of the roll of oilcloth.

As you might guess, providing a method of estimating this quality or otherwise developing a basis for decision involves statistics.

These three examples come from three different sectors of the business world. However, they all have various things in common. All problems are characterized by alternatives. They all call for decision (the selection of one alternative) and action. In addition, they all can be treated by statistical techniques. As noted, these techniques will be studied in due time. Don't be impatient.

1.4 States of the World

In any problem certain alternatives are good. Some are bad. Some are better than others. Some are worse than others. But what do we mean by good, bad, better, or worse? Simply put for the moment, our primary task in the next few paragraphs will be to try to give some meaning to the phrases *good alternative* or *decision* and *bad alternative* or *decision*. This is obviously a basic consideration, for in order to adopt rules for decision making we must have some sort of notion of why and when certain decisions are desirable.

To be able to classify alternative courses of action as good or poor or inferior or superior or as something else, we must think in terms of the *consequences* of a given decision. For example, in any problem, if one of the decisions open to management will *always* yield better results than any other decision no matter what happens, a wise course of action (and an obvious one) is to make that decision. We might say, that that particular alternative dominates all others. Then, no other course of action need be considered seriously.

More often a different and more difficult situation exists. Under

certain circumstances one decision may be preferred to a second since the consequences associated with the first are more desirable. However, under different conditions the first decision may be inferior to the second since the consequences associated with the second are more desirable. Thus, decision A may be better than decision B, or decision B may be better than decision A, *depending upon certain conditions*.

Consider the following nonstatistical but homey illustration:

Example 1.6 Recently one of the authors had to visit a dentist whom he had not had the pleasure nor the pain of visiting previously. He was told the office was at 654 Madison Avenue, New York City—somewhere in the 60's; that is, between 60th Street and 69th Street.

He decided to ride the subway but absent-mindedly forgot to get more detailed directions. However, he did realize that the subway stopped at 59th Street and at 68th Street, among other places.

He faced a decision problem with incomplete information. His alternatives were to get off the train at 59th Street and to get off the train at 68th Street. Which was a better alternative? This depended upon the location of the dentist's office. If the office were at 62nd Street, a decision to get off the train at 59th Street would be preferable; if it were at 67th Street, the 68th Street stop would be better; and so forth.

The desirability of a particular decision depended upon a certain condition—the (unknown) location of the dentist's office. In the absence of the information about this pertinent fact, the prospective patient had to make his decision in the face of uncertainty about which alternative was the better.

To be able to say with certainty which alternative course of action is best in a given decision problem, we ordinarily must know certain conditions about the situation we face. What are these conditions? In general, we will refer to them as the *true state of the world*. A true state of the world is a complete and accurate description of a given problem situation. Note that in most problems there are many *possible* states of the world. Therefore, a basic question in decision making is which of these many states is actually the true one. Ordinarily, it is impossible to answer this question exactly—a matter we shall consider in the next section. First, consider two more examples.

Example 1.7 A personnel administrator must decide whether or not to hire an applicant for a job. One way of defining states of the world in the problem is (1) potential success, and (2) potential failure. Each of these states of the world describes the applicant's characteristics which are pertinent to the decision problem.

Example 1.8 An accountant wishes to value the end-of-year inventory of a wholesaler. States of the world are represented by possible values of the inventory. One state of the world is \$52,318. Another is \$26,321.15. Still others are \$121,300.01, \$2,000, and \$58,111.58. Obviously there are quite a few possible states of the world.

We can summarize some of our earlier remarks through the use of this new terminology. The desirability of alternatives depends upon which state of the world is true. Thus, if a decision maker knows which state of the world is true in a given problem, he should have no trouble in making a decision. For example, if the accountant of the last illustration knew that the true state of the world (the true value of the inventory) was \$55,090.10, his decision on what value to attach to the inventory would be straightforward—undoubtedly it would be \$55,090.10. If the personnel administrator of the next-to-last example knew the state of the applicant, he would know whether to hire him or not hire him.

Consequences refer to the results of making certain decisions when various states of the world are true. They are the outcomes of making a decision. Any outcome depends upon which alternative course of action is selected and upon which state of the world is true.

Example 1.9 A manufacturer produces a certain type of fuse, whose quality he measures by whether or not it blows out at 20 amperes. The manufacturer usually sells his product in lots of about 1,000 fuses, and he has the following alternative courses of action: (1) sell a lot as *first grade* for \$100 (10¢ a fuse) with a double-your-money-back guarantee for each fuse that proves defective, and (2) sell a lot as *second grade* for \$85 (8½¢ a fuse) without any guarantee.

States of the world may be described by the number of defectives in a lot.

If he adopts alternative (1), the manufacturer will make \$100 per lot less 20¢ per defective fuse (assuming that all defectives are returned under the guarantee). However, if he adopts alternative (2), the manufacturer will make a straight \$85 per lot.

Thus, the outcomes for a few of the many possible states of the world and for the two alternative courses of action may be summarized as follows:

Alternative	State of the World (Number of Defective Fuses)					
	0	10	25	50	75	100
Sell as first grade.....	\$100	\$98	\$95	\$90	\$85	\$80
Sell as second grade.....	85	85	85	85	85	85

Thus, for any given lot, if the manufacturer knew that only 10 fuses were defective, he would choose alternative (1) and make \$100 less \$2.00 (20¢ times 10), or \$98. This is the desirable alternative since under alternative (2) he would make only \$85.

On the other hand, if he knew that 100 fuses were defective, he would choose alternative (2) and take the \$85 since under alternative (1) he would make only \$80 (\$100 less 100 times 20¢).

In general, the student should satisfy himself that if less than 75 defectives are in the lot, it should be sold under alternative (1), while if more than 75 defectives are in a lot, it should be sold under alternative (2). If exactly 75 defectives are in a lot, neither alternative is preferable to the other.

One last word on this problem. We have assumed that the gross income from a lot adequately measures the consequences of a decision. The validity of our discussion is, therefore, based on the premise that no damage to the manufacturer's reputation occurs because of returned fuses. If this premise is not met, our sketch and analysis of the problem would be somewhat inadequate.

In many problems it is quite difficult to measure consequences numerically. Often we can only evaluate them qualitatively.

Example 1.10 The personnel administrator we met before (in Example 1.7) considers the consequences of hiring or not hiring an applicant. If the applicant is hired and he is a potential success, the firm would be rewarded by having a competent employee. If the candidate is hired and he is a potential failure, the firm would be wasting time and money on his training. In addition, the hiring of a noncompetent person might have an adverse affect on the morale of other employees and might result in incalculable damage to customer relations.

In either event, you will observe, the consequences of the action of hiring depend on the true state of the world. Further, the consequences in this situation cannot easily be converted into dollars and cents but must be expressed qualitatively in terms of advantages or disadvantages to the firm.

Similarly, if the action is not to hire the candidate and he is a potential success, the firm, as a consequence, would have to forego the services of a capable person. It might have to leave an important position unfilled and thus interfere with the progress of an important program. On the other hand, if the candidate is a potential failure, then the consequence of the action—not to hire—would save the firm the trouble and expense of trying to train a noncompetent person. Again the consequences are related to the action taken and the true state of the world; and again the consequences are difficult to measure numerically.

However, the thought processes of the personnel administrator in analyzing the consequences of his alternate actions may be summarized qualitatively as follows:

Alternative	State of the World	
	Potential Success	Potential Failure
Hire.....	Desirable consequences	Undesirable consequences
Not hire.....	Undesirable consequences	Desirable consequences

In short, the first step towards the solution of any decision problem is to spell out the alternative courses of action, the states of the world, and the consequences associated with each possible outcome. This is a specific way of stating a common-sense rule—the first step towards the solution of any decision problem is to recognize what problem exists.

1.5 Uncertainty, Observations, and Sampling

Whenever a decision maker knows which state of the world is true, he should be able to make a rational decision since a state of the world is supposed to be a complete description of the pertinent facts in a problem situation.

Unfortunately, many business decisions must be made without knowing the true state of the world. They must be made in the face of incomplete information about which state is actually true. These problems are decided under uncertain conditions. A manufacturer cannot know the number of defective fuses in every lot upon which he must make a decision. A personnel administrator cannot know with certainty whether each applicant coming before him is a potential success or a potential failure. So it is in many other types of decision problems: there exists some degree of *uncertainty* about the true state of the world.

The objective of using statistics is to recommend action in the face of this lack of complete information or in the face of this uncertainty. Put another way, one of the functions of statistics is to provide information on which state of the world is true. This reduces the amount of the uncertainty that exists.

And put still another way, the objective of statistics is to provide *observations* or numerical information to be used as the basis for decision. This ordinarily does *not* mean that statistics involves counting the number of errors in *all* paper work, or counting the number of defectives in *all* product, or measuring the diameters of *all* ball bearings in a lot, or asking *all* television viewers their favorite programs. These usually are expensive or otherwise undesirable ways

of providing the required information. Rather, one may utilize the theory and the methods of statistics so that he may get the necessary information in the quickest, cheapest, and most efficient way.

It must always be remembered that the acquisition of observations or data costs money, time, and effort. We must, therefore, always weigh the cost of added information against the gains derived from reducing the amount of uncertainty. It would obviously not be wise to spend \$1,000 to gather information that might save \$100 or yield an added income of \$100. We must always try to find the most economical method of reducing uncertainty. For this and other reasons we will consider ultimately, statisticians often find it desirable to provide only partial or *sample* rather than complete information.

Example 1.11 In an earlier illustration, a manufacturer of fuses had to decide for each of his lots of product whether to sell under a guarantee or without a guarantee. One statistical solution to this problem might be for him to take a sample of 40 fuses from each lot. He might then utilize this sample as a basis for action on the lot by testing the 40 fuses and selling the rest of the lot under the guarantee if and only if there are less than 3 defectives in the sample. We will return to this type of problem for the important details in a later chapter; for the present you should merely consider the idea of sampling.

Decision making based on sample information may seem like a risky business. That it is. Any time decisions are made on the basis of incomplete information there are risks of making an incorrect decision—a decision which would not have been made if complete information had been available.

Example 1.12 The manufacturer of fuses, last noted one paragraph ago, runs some risks if he adopts a sampling plan such as the one sketched in the last illustration.

There is a risk that if by chance the sample is poor, he will sell the lot for the lower price without the guarantee, although actually it is of good quality and should be sold under a guarantee.

On the other hand, there is the risk that if by chance the sample is good, he will sell the lot with a guarantee, although actually it is of lower quality and should be sold without a guarantee.

Despite the risks due to sampling, the manufacturer cannot get complete information by testing every fuse. Why?

Example 1.13 A production manager in a plastics company must estimate the total number of widgets he should produce next week. He knows some widgets will be defective, so to produce 10,000 good widgets he must plan a total output of over 10,000.

His alternatives are 10,000, 10,001, 10,002, . . . , 11,000, . . .

12,000 . . . There are many risks of incorrect decisions in this problem. For example, one incorrect decision is producing 10,000 widgets when 10,500 will be needed. Another is producing 12,000 when 11,111 will be needed.

Note, however, some incorrect decisions are not as bad as others. This is judged on the basis of consequences, the costs of overproducing and of underproducing. Thus, overproducing by 10 items ordinarily would not be as bad as underproducing by 200 items. Certainly overproducing by 10 items is not as bad as overproducing by 100 items.

Also keep in mind this bit of philosophy: The more costly the consequences of an incorrect decision, the smaller the *risk* one should be willing to take of making that incorrect decision.

Despite the risks of incorrect decisions whenever sample rather than complete information is obtained, sampling often is preferable. Savings in money and in time are but two of the reasons why one might prefer to collect partial data. A statistician recognizes these advantages but at the same time realizes the risks. That is his job.

In most problems, therefore, he insists on a particular type of sampling which is termed "probability sampling." Samples are drawn or considered in terms of the mathematical theory of probability, a procedure which allows a decision maker to measure the risks inherent in the process of making decisions based upon incomplete or sample information. This whole area represents another central theme which we will meet again and again in our studies.

It should be remembered from the outset of our studies, however, that in some problems a statistician will not recommend the use of a sample. Rather he may say: "Make your decision without the benefit of observation. It is not feasible to attempt to get any information—it will yield less than its cost." Or, in the other extreme type of problem, he may say: "Utilize a complete count and not a sample for this problem—it will be more economical."

In conclusion it might be stated that one uses statistics to set up a procedure for obtaining and interpreting observations. However, any such plan of attack must be centered around the task of solving the business problem on which he is working. And remember that by the term "solving" we mean affording management a rational basis for decision.

1.6 Testing and Estimation Problems

At various points throughout this text we shall classify and define different types of statistical studies. This section presents the first of these classifications. It deals with the difference between statisti-

cal *testing* problems and statistical *estimation* problems. These two types of studies both have the same general purpose: providing a basis for decision and action on a business problem. The distinction is important, however, in the design of a study—different plans must be made for testing problems than for estimation problems, and vice versa.

The feature that distinguishes testing problems from estimation problems is the nature of the alternatives available to a decision maker. If the alternatives are nonnumerical or qualitative, the problem may be regarded as a testing problem. On the other hand, if the alternatives are numerical (quantitative, that is), the problem may be regarded as an estimation problem.

Thus, testing problems are characterized by the fact that management has two or more qualitative alternative courses of action and decision available. The purpose of a statistical study is to decide which course of action appears to be better *on the basis of numerical observations*. We have had several illustrations of testing problems, the last of which was contained in Example 1.12. There the decision problem was to classify a lot of fuses as first grade or as second grade and accordingly to sell the lot with or without a guarantee. The alternatives are nonnumerical; therefore, the problem is treated as a testing problem.

Estimation problems are characterized by the fact that management has quantitative alternatives available. The purpose of a statistical study is to recommend one of these *numbers* for use. Put another way, the purpose of a study in an estimation problem is to estimate some quantity. Example 1.13 dealt with an estimation problem—estimating the level of production. For another illustration consider the following:

Example 1.14 A small company manufactures various electronic tubes, including a special type of transmitter tube. At the end of the fiscal year the company accountant must find the approximate value of the tubes in inventory. The number of tubes on hand is known, but the accountant realizes that some tubes are not saleable because they are broken or defective. He concludes, obviously, that these tubes should not be valued in the inventory.

In any event, the accountant's alternatives in valuing the inventory are numerical, being typified by such numbers as \$10,125, \$235, and \$321,123. The problem is an estimation problem.

Incidentally, a statistical solution to this problem would be to base an estimate of the value of the inventory on a sample of transmitter tubes. If this solution were implemented, what risks would the accountant be taking?

Once again, remember that whether they are used in testing or in estimation problems, statistical theory and statistical methods are decision-making tools. This follows immediately from the thought that both testing and estimation problems are types of decision problems.

1.7 The Principle of Planning

A decision problem exists when there are two or more alternative courses of action. Decision is a summary word for the process of selecting one of these alternatives. The desirability of any decision is measured by the consequences of that decision. These consequences depend not only on the action decided upon but also on which state of the world is true.

Many, if not most, business decisions must be made in the face of uncertainty. This means a decision maker is not sure which state of the world is true. This, in turn, means that he is not sure of what consequences will result from any decision he might make. Hence, he is uncertain about which decision to make.

Uncertainty arises when a decision maker has incomplete information about which state of the world is true. Statistics is a field of study in which there exist theory and methods for reducing this uncertainty, within economic limits, by providing sample information.

Sample information will not remove all the uncertainty from a decision problem. There will still be some uncertainty, hence some risks of incorrect decisions. But with the use of statistical theory and statistical methods, these risks can be measured and controlled.

Because of the fact that incomplete information is the rule rather than the exception, too often the gathering of information becomes the thought-consuming problem to a decision maker. Too often the decision problem thus simply reverts to a data-gathering problem. Too often little thought is given to how data will be used once they are obtained. The unfortunate result is that too often data are gathered but, when available, they furnish no guide to action.

Example 1.15 A consulting firm was retained by a municipality to assist in the relocation of residents whose dwellings were to be condemned as part of a slum-clearance program. The municipal officials stressed their desire to carry out the relocation program with a minimum of disturbance. They felt that social and economic data about the residents were of the utmost importance in attaining this objective.

Armed with this advice, the consulting group made two visits to

the area. It then seriously considered whether or not to set forth upon a third data-gathering expedition.

At this stage, one of the authors was invited to enter the picture. He naïvely inquired if the consulting firm had planned what to do with the data after they were acquired. How would they aid in the relocation of families? He was surprised to hear that a plan for utilizing the data had not been formulated. Unfortunately, the municipal officials in presenting the problem had put such great emphasis on the paucity of data that the collection of data had become an end in itself. The failure to do a little a priori planning had resulted in much wasted motion and cost.

Alternate programs of relocation should have been worked out beforehand. The data necessary for the selection of an alternative could then be gathered. In any event, the data needed to reach a decision would have been more readily apparent.

To gather data that will not help one make a decision is wasteful of time and money. In addition, it is alien to common sense. It is also alien to the field of statistics. Most statisticians view the proper function of their subject as relating decisions to the data one plans to collect *before* the data are actually collected. This philosophy calls for a decision maker to quiz himself as follows: "What will I do with these data?" "What decision will I make if I observe a certain result?"

Statistics has in recent years developed to the point where the emphasis is not on data but on *design for decision*—designing a study so that it does more than furnish facts—so that it directly leads to decisions. This point of view, therefore, stresses a priori planning.

This idea can be implemented in many studies by incorporating a *statistical decision rule* into the plan of attack. In such a study the decision maker formulates his possible decisions a priori, that is, before the actual observations are made. However, these a priori decisions are conditional upon what he observes. A simple form of a statistical decision rule is: "Take a sample of such a size in such a way; if we observe so-and-so in our sample, then we will do this; otherwise we will do that."

To be more specific about such, so-and-so, otherwise, this and that, let's look at two illustrations covering a personal decision problem and a personnel decision problem in that order.

Example 1.16 A homey example of an a priori decision rule was recently implemented by one of the authors and his wife. Expectant parents, they faced the decision problem of naming their expected arrival. They made their decision a priori, relating the action to be

taken to an observation which was to be made subsequently—the sex of the child. Their rule was as follows: If the baby is a boy, name him Louis; if the baby is a girl, name her Laura.

Their decision rule was somewhat incomplete, but only in the sense that they did not provide an action for twins, triplets, etc. Fortunately, for more reasons than one, their rule did suffice.

Of course, one need not be a statistician to use this type of rule. Thousands of expectant parents have used a similar plan—it is quite commonplace, to say the least.

Example 1.17 A personnel administrator may use the following decision rule in deciding whether or not to hire secretarial help: If the girl can type at least 60 words per minute and take shorthand at the rate of at least 100 words per minute and score over 75 on a given aptitude test, hire her; otherwise don't hire her.

Note in particular that the rule relates decisions to observations before the observations are made, as all good decision rules must.

Throughout this book we will be studying the theory and the methods of constructing rational decision rules. The word “rational” is important in this sentence—it means that we will be studying ways of constructing rules to achieve certain managerial objectives. We will learn more about this as we go along in our studies.

1.8 The Principle of Teamwork

Planning is a basic principle of statistical applications. Who is responsible for this planning, and how is it achieved? Briefly, planning is a team function and may be achieved only through teamwork.

Most business problems are solved not by one man but by a group of men working in concert. Such a group, insofar as they have identical interests in mind, may be thought of as a team.

Depending on the nature of the decision problem, the general manager, the production expert, the marketing man, the sales manager, the advertising director, the controller, the internal auditor, and/or the credit manager may be important players on a given team. Throughout this book, we will take the liberty of grouping all these players together and, for convenience, call them *management*. Management is that group which directs the strategies and makes the decisions of a firm in the game called business.

In some problems—those that are statistical—a statistician is a member of the team. He performs a useful function. In this text, we will, of course, be interested primarily in what position the statistician plays on this team and what he can contribute to the play-

ing of the game of business. We will often refer to management and the statistician because our interest is primarily in statistics. But this should not imply that the latter is a separate entity or the individual star. A statistician works with others who direct the team's strategies, and he works for the team. Remember, the statistician is an important part of a business team, but he is important only as part of that team.

Sometimes one player may act as a statistician and also as, say, a marketing expert. Sometimes, in fact, one player may be the whole team. In any event, we are going to study the statistical part of the teamwork that is required to solve many business problems—to make decisions in the face of uncertainty on the basis of numerical information.

We have stressed the idea of a team because co-operation is the key to the planning of any statistical study. Many times the nature of a business problem may, at first, be somewhat vague. The subject matter experts on the business team and the statistician must then sit down for a conference and discuss in detail the nature of the problem. The alternatives must be listed, and the possible consequences of each alternative thought out. States of the world must be defined. Given a clearer view of the whole problem, the team must then decide if the problem is statistical. If so, then they must also decide upon the appropriate statistical framework. This involves other elements which the team must formulate, and to which we shall turn in Chapters 2 and 3.

1.9 You Should Now Know That

Statistics are data, but, in addition, statistics is a field of knowledge.

The field of statistics involves methods and theory as they are applied to numerical data or observations with the objective of making rational decisions in the face of uncertainty.

Statistics provides a basis for action in business decision problems.

A decision problem exists when two or more alternative courses of action are available. In many problems statistics provides an objective basis for selecting one of these alternatives.

A state of the world is a complete description of the pertinent facts surrounding a given problem situation.

Consequences are the results of making a certain decision when a certain state of the world is true.

In many problems a decision maker may be uncertain about which state of the world is true, hence uncertain about which decision to make.

Statistics provides methods of making sample observations which reduce the uncertainty connected with business problems.

Statistical studies deal with either testing problems or estimation problems.

The proper function of statistics is to relate decisions to the data one plans to collect *before* the data are actually collected. This is the principle of planning.

Statistical studies usually must be formulated by management *and* the statistician. This is the principle of teamwork.

1.10 You Should Now Be Able to Solve

1. Some states require a minimum bodily injury liability coverage on automobiles licensed in their state. Many automobile owners, however, decide to carry as much as 10 times the amount of minimum insurance.

Discuss the nature of the reasoning that results in the decisions of these individuals to carry large amounts of insurance. Frame your answer in terms of alternatives, states of the world, uncertainty, consequences, and numerical information.

2. You have been accepted for admission to three graduate schools to which you have applied. You have a month within which to choose one, or none, of them. Is this a statistical decision problem? Discuss the problem in terms of alternatives, states of the world, action, numerical data, consequences, and uncertainty.

3. Discuss the adage "don't cross your bridges until you come to them" generally in terms of statistical decision theory and particularly in terms of the principle of planning.

4. Every month the Department of Commerce conducts a sample survey of retail sales in the United States. Retail sales for March, 1958, were reported to be \$15.6 billion.

- a) Is this an illustration of a testing or of an estimation problem? Explain.
- b) What are the alternative courses of action in such a problem?
- c) Is there any uncertainty in this decision problem? Why? Does the sample survey eliminate the uncertainty? Why or why not?

5. You must decide between going to the movies with a friend tonight and studying for a statistics quiz on Chapter 1 that may be

given tomorrow. Consider this problem in terms of the concepts and principles of decision theory.

Make a decision and explain the reasoning process which you used in making the decision.

6. A department store's management wants to study, on a day-to-day basis, the records and receipts of the clerks responsible for refunds. The internal auditor asks what the purpose of the study will be (will they retain or fire certain girls or maintain or change the refund system, etc.). Top management says, "Let's do the study first, and then we'll formulate a course of action." Is there anything wrong with this answer? Why or why not?

7. Select some decision problem with which you are familiar from your own field of specialization. Discuss in terms of alternatives, states of the world, consequences, uncertainty, etc., and in terms of how you think statistics might help in arriving at a decision in the problem.

8. An investor in the stock market must decide whether to hold or to sell his portfolio of stocks. Discuss this situation as a decision problem.

1.11 You Will Also Find That

The ideas of decision, alternatives, action, *et. al.*, are covered lucidly in:

BROSS, IRWIN D. J. *Design for Decision*, pp. 1-32. New York: Macmillan Co., 1953.

Decisions in the face of uncertainty, states of the world, consequences, and related topics are discussed very clearly in:

SAVAGE, LEONARD J. *The Foundations of Statistics*, pp. 6-10, 13-17. New York: John Wiley & Sons, Inc., 1954.

Some further examples and remarks about the nature of statistics and statistical planning are given in:

SPROWLS, R. CLAY. *Elementary Statistics*, pp. 1-8. New York: McGraw-Hill Book Co., Inc., 1955.

For some basic thoughts on what statistics is and/or what statistics are, the following are key references:

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics: A New Approach*, pp. 3-16. Glencoe, Ill.: The Free Press, 1956.

MCCARTHY, PHILLIP J. *Introduction to Statistical Reasoning*, pp. 1-6. New York: McGraw-Hill Book Co., Inc., 1957.

CHAPTER • 2

Statistical Populations

2.1 Defining the Problem Statistically

Our first chapter was devoted to a discussion of some of the aspects and principles of decision making. Many of the ideas are rather general—they apply to a variety of business and personal problems. Naturally, there are many such problems. There are also many ways of arriving at decisions.

As was pointed out, one of the ways of arriving at decisions in some problems is through the effective use of statistical methods and statistical theory. Thus, if any problem is such that there is uncertainty about the true state of the world and numerical data will help reduce this uncertainty, then that situation presents, at least in part, a statistical decision problem. And if one is faced with a statistical decision problem, the next step in the decision-making process is for management and the statistician to translate it into statistical terms.

Incidentally, this translation is not wholly a statistical job. It is a team job. It requires the combined efforts of the subject matter experts and the statistician who are working on the given problem.

The statistical definition of a decision problem usually involves two integral parts. If you will tolerate the use of technical terms for a moment, we may call these two things:

1. Defining the statistical *population*, and
2. Specifying the *decision-parameter*.

This chapter is devoted to developing the idea of defining the statistical population. In statistics the term “population” is a special word with a special meaning—you should immediately forget the conventional meaning you attach to this word. In its place, re-

member that a *population* is the totality of all pertinent observations that might be made in a given problem. You might also make note of the fact that an often-used synonym for population is *universe*. On the other hand, if only a part of a population is selected in a study, that part is called a *sample*.

The second important part of the statistical definition of a problem, specifying the decision-parameter, will be considered in some detail in Chapter 3. As you will see, it involves methods of describing the state of the world in a problem; that is, it involves ways of statistically summarizing the pertinent facts about a population.

Example 2.1 A marketing research study is planned to provide information for a decision as to the amount of money a company will spend for a local advertising campaign to promote its oleomargarine. Included in the type of data desired is the amount of oleomargarine used by the families in the specified area.

The statistical population consists of the different amounts of oleomargarine the various families use. If the population could be made known, it might be written (in terms of the number of pounds per week) as 0, 1.5, 0, 1, .25, 0, etc., where each number is associated with a given family.

Thus, the population for this problem is a set of numbers—each number is a possible observation and one bit of information. The totality of these numbers, or observations, or bits of information, is the population.

In practice, it might be costly and time consuming to study the population completely. However, some information (from a sample of the families) might be obtained by the use of statistical methods and a decision made on the basis of the observations in the sample.

A decision-parameter, incidentally, is some characteristic of this population; for example, the average amount of oleomargarine used by all the families, or perhaps the proportion of nonusers. But, as noted, let's save this for the next chapter.

Although we now are going to spend a full chapter on populations, statistics deals primarily with the making of decisions on the basis of samples. Complete or population information is often unobtainable. Although there are exceptions, usually one must settle for sample or partial data as a basis for decision making.

Despite these facts, it is worthwhile, at this stage of our studies, to ask ourselves what we would do if complete information—the population—were available or could be made available. It is worthwhile because this is a question which management and the statistician must invariably ask in the planning stage of any proposed statistical study.

First of all, an answer to this question points up clearly what type of information is needed. Whether this information can be obtained is another matter. After all, before thinking about whether you can get certain information, you should think about what information you need to make the decision that you are facing. Secondly, thinking about complete information focuses attention on the decision problem, *per se*. It directs attention to possible actions and not to the gathering of data which may turn out to be useless when collected.

Thus, thinking about complete information in the planning stage of a study is a fruitful procedure. It emphasizes what we have called *a priori* planning, one of the cardinal principles of statistics.

2.2 Elementary Units and Their Characteristics

Let's examine the whole idea of observations in more detail. Many statistical studies involve observation of *characteristics* of what may be called *elementary units*. For example, one decision problem may involve observations of the quality of fuses. Thus, an elementary unit is a fuse; the characteristic to be observed is its quality. Alternately, a problem may require observations of the net income of corporations in 1958. In this case, an elementary unit is a corporation; the characteristic to be observed is net income in 1958.

Example 2.2 To extend our introduction to the present topic, we may consider the following list which presents a variety of elementary units and some of their characteristics which may arise in various statistical decision problems:

<i>Elementary Unit</i>	<i>Characteristic</i>
An apartment.....	Availability
A dwelling unit.....	Type of heating
A family unit.....	Annual income
A student.....	Quiz grade
A person.....	Marital status
A firm.....	Net profit per sales dollar
An account receivable.....	Amount due
A common stock.....	Dividend rate
A product.....	Price
A worker.....	Employment status
An employee.....	Absences in one month
A bank.....	Number of depositors
A radio tube.....	Time to failure
A steel rod.....	Tensile strength
A steel rod.....	Diameter
A steel rod.....	Weight

We have already defined a statistical population as the total set of observations that may be made in a study. In our new terminology,

you may also think of a population as the complete set of elementary units. In any event, remember that those observations actually made in most statistical studies are only a sample—a part, but not all, of the population.

Example 2.3 The marketing research study planned to provide a basis for action on an advertising campaign in Example 2.1 involved the amount of oleomargarine used by families located in a certain area.

Note that:

1. The elementary units in this study are the families in the specified area.
2. The characteristic of interest for each family is the amount of oleomargarine it uses in a given week.
3. The whole set of observations of this characteristic is the population under study.
4. Generally a statistical study would involve observing only a sample. Why?

Example 2.4 A company accountant must decide whether or not to check this week's payroll vouchers for any numerical or other mistakes that may have occurred.

Note that:

1. In such a problem, the elementary units are all the payroll vouchers under consideration.
2. The characteristic of interest for each of these units is its quality. This, in most applications, may be defined as either "in error" or "correct."
3. Once again, the total set of these observations is the population.
4. Decisions would ordinarily be based on a sample of these observations.

Example 2.5 A manufacturer receives a large quantity of a special machine bolt that he uses in his production process. He must decide whether or not to accept the lot. He can only use a bolt if its diameter is between .2405 inches and .2415 inches.

Note that:

1. An elementary unit in this study is a bolt.
2. The characteristic of interest for each bolt is its diameter.
3. The whole set of diameters is the population.
4. Usually, however, a decision will be based on a sample of the diameters. Why?

Also note that an alternative definition of the characteristic of an elementary unit is available for this problem. Those bolts between .2405 inches and .2415 inches might be classified as "good" and others as "defective." Then this set of observations is the population, and a sample of these observations may be used as a basis for decision.

The last illustration points up the fact that in a given problem there may be different ways of specifying the population, simply because there may be different ways of observing characteristics. However, while the choice of an appropriate characteristic, and the choice of the appropriate elementary unit as well, depend upon many factors, they usually follow from the nature of the decision problem. Nevertheless, it should be stressed that these choices must be made early in the planning stage of a study by the subject matter experts and the statistician working in concert.

Example 2.6 A household appliance company wishes to determine, within a certain area, the number of dwelling units in which there are electric refrigerators, in which there are gas refrigerators, etc., so that they may better plan their sales campaign in that area.

A sample study of dwelling units is considered. Thus, if the study is to be made, it seems logical to define a dwelling unit as the elementary unit. This is a start—but we must be more specific about the definition of the dwelling units. Should apartments be included? What about hotel rooms with kitchen facilities? Should lodging or boarding houses be included?

Answers to these questions must be provided in the planning stage of the study. The elementary unit must be defined very specifically before the survey itself is even begun.

There is an important lesson to learn from this illustration. Although the definition of the elementary unit is an important consideration in a statistical study, it is not exclusively a statistical problem. Its solution depends, of course, upon what type of information is wanted by the company, and this must be clearly described by the marketing, advertising, and sales experts who are working on the problem.

Example 2.7 A common problem in production control involves deciding whether to let a machine continue to operate as it is or to stop the machine and reset it. It is also commonplace for such decisions to be based on a sample of product taken from that already produced by the machine. What should be observed by, for example, the shop foreman when he samples the product? The elementary unit is a part, or a can, or a fuse—whatever product is being produced. But which characteristic of the product is of interest?

If the product is a steel rod, such quality characteristics as breaking strength, diameter, weight, or length might be observed. If the product is a can of fish, the weight, the aroma, the flavor, or the appearance of the fish might be checked. Should all of these characteristics be observed? Or, if not, how are the characteristics of interest to be selected?

A general rule is to observe those characteristics which, in the past, have led to the most expensive and serious losses when they were not being produced according to specifications. Whatever characteristics are decided upon, however, they must be decided upon a priori—in the planning stage of a study.

As these last two examples indicate, the selection of an elementary unit and of the appropriate characteristic is important to point a study towards a good solution. Remember, the aims of statistical studies generally involve finding out something about the characteristics of elementary units. Why? In order to make a rational decision despite some uncertainty. Unless we know the elementary units and the characteristics in which we are interested, a study may very well be a waste of time and money.

2.3 The Frame

In the last section we considered two important points that invariably must be dealt with when one defines a population. How is an elementary unit to be defined? What characteristic is to be observed? Answers to these questions are necessary before a study can proceed. In addition, those in charge of planning the study must then consider whether or not their definition of the population is operationally feasible; whether or not there exists any way of access to the population. Having considered where they want to go, they may need, so to speak, a “road map.” In statistics such a road map is called a frame.

Access to the population is no problem in many applications of statistics. Thus, when you must decide whether or not to accept a lot of product, the lot is usually before you, more or less accessible and available for study. In other cases, however, access to the population presents much more difficult problems. For example, it is easy to define a population as the brands of all radios in use. But where are all these radios? Can we identify their locations in terms of a frame? That is now the important question.

We may formally define a *frame* as a list or map or some other concrete way of identifying all of the elementary units to be covered by a statistical study. Such a “list” is quite necessary in order to carry out many studies. Unless we have identified where and from whom we can gather the information we desire, how can we possibly obtain it?

Example 2.8 A controller must analyze his firm’s accounting operations over the past several years to derive certain financial infor-

mation not ordinarily tabulated; e.g., the number of checks drawn for amounts under \$2.00. The analysis is to be the basis for a decision on whether or not the company should revise parts of its accounting system.

Part of the study can be carried out through an analysis of canceled checks. Thus, an elementary unit is a canceled check. One characteristic of interest is whether or not the amount of the check is less than \$2.00. It is noted that the canceled checks have been filed by number; and that there are up to 1,250 checks per file drawer. A frame for the proposed study is easily constructed to cover all checks from July 1, 1956, to June 30, 1958 (the dates set for the study). It may be written down quite simply, as follows:

File Drawer Number	Includes Checks Numbered	Number of Elementary Units
112.....	4086-5000	915
113.....	5001-6250	1,250
114.....	6251-7500	1,250
115.....	7501-7624	124
		3,539

This frame, you should note, identifies each and every elementary unit—all 3,539 checks—from July 1, 1956, through June 30, 1958. It therefore shows the location of all the elementary units of interest, thereby operationally defining the statistical population and providing a basis for implementing the proposed study.

Example 2.9 An accountant must value, at retail, the end-of-year inventory in a retail shoe store for tax purposes. He defines a pair of shoes as an elementary unit; the retail price of a pair of shoes as the characteristic of interest. Then a frame may be developed in terms of a series of diagrams, each diagram identifying shoe boxes or pairs of shoes in a certain section of the store.

Example 2.10 The Turkish government conducts a quinquennial census of population. In many Turkish villages, however, there are no streets and houses have no numbers. How can frames be constructed for the study? One solution, used prior to 1955, was that every mayor had the responsibility of numbering every house and of preparing a list of the numbers identifying their location. These lists, then, represented frames of dwelling units in the villages.

It is important to remember that a frame is essential for a census (complete count) as well as for a sample study. The only difference in the two situations is that for a census we make observations on *all* the units included in the frame, while for a sample we observe only some well-selected portion of the units. Thus, to conduct a complete

or a sample study of faculty members in a university, there must be a frame to specify and to identify the teachers to be included. Otherwise a survey would be only a hit-or-miss affair. This same idea holds true for every study.

Many times a frame is readily available, or at least readily constructed, for a statistical study. However, occasionally it is difficult to define a satisfactory frame—particularly for statistical studies carried out over a wide geographical area. This is one basic difficulty in marketing and advertising research covering a regional or a national market.

Some common sources of frames for “widespread” surveys are telephone directories, customer lists, lists of voters, automobile registrations, tax assessor’s lists, and city directories. However, with the possible exception of the last two, such frames are adequate only for limited purposes. Usually, they are not satisfactory for statistical studies of the general public. Therefore, extensive use often is made of geographic areas to identify *groups* of elementary units rather than individual elementary units. The frame in such a case may be a series of maps of geographic areas. Such frames include detailed maps of counties, standard metropolitan areas, census tracts, city blocks, and census enumeration districts. Each of these areas may serve as a way of identifying, or at least locating, dwelling units, places of business, or consumers.

Frames are subject to various imperfections, and, in practice, it is very often difficult to avoid all of these potential sources of trouble. It is essential, therefore, to determine at the outset of a statistical study whether a given frame is satisfactory.

For one thing, a frame is *incomplete* if it does not include every elementary unit that should be included within the scope of a study. We may or may not be aware of the fact that a frame is incomplete. This imperfection ordinarily will be more serious if we are not aware of it. For one thing, incompleteness in a frame is unlikely to be discovered in the course of a study. For another, it can be a serious defect if the incompleteness is confined to units possessing a special characteristic, for these units would, as a consequence, be under-represented in a study.

At least when we are aware of incompleteness we can decide whether or not it will invalidate the study, and act accordingly. You see, at times it is not possible to construct a frame to include all elementary units, and a frame with some units omitted must be considered. We may designate the population represented by an in-

complete frame as the “working” population as opposed to the population we are interested in—that is, the “target” population. It is important to recognize that any survey results, if and when obtained, will refer only to the working population. The key question is then: “Is this information *adequate*?”

Example 2.11 A department store considers expanding its T.V. service department and offering, among other things, a new one-year service contract. How many new customers and what type of sets (new or dilapidated) will the service contract offer attract? Answers to these questions will provide a basis for deciding the extent of the department's expansion and the rate to be charged for the service contract.

Management desires to test the offer on a hundred or so customers before presenting it to the public at large. From what group should the sample be selected? What is the population? Is it represented by all T.V. set owners in the New York metropolitan area? If so, a complete frame would be almost impossible to obtain.

However, suppose that the store does have a list of persons who have purchased their sets there and lists of other persons who have employed the store's service department as well. Does this provide a satisfactory frame even though it restricts the scope of the study? This is a question which management and the statistician must deliberate very carefully in the planning stage of the proposed study.

As contrasted to incompleteness, a frame also may be imperfect because it contains duplicates of certain elementary units. A study based on that frame may be biased. Of course, when duplication is present, it often can be corrected by an examination of the frame, although such an examination may be a tedious and costly process.

The success or failure of many statistical studies depends on the type of frame that is available. For one thing, the definition of the population is tied closely to the frame selected. Therefore, it is unwise to plan a study in any detail until one has knowledge of the availability and of the accuracy of a frame. Needless to say, if a reasonable and adequate working population, defined in terms of a frame, cannot be found, then a proposed study must be discontinued.

2.4 Enumerative and Analytical Studies

A statistical population, as we have defined it, consists of a set of observations pertinent to a decision problem. These observations may be finite or infinite in number, depending upon whether a statistical study is *enumerative* or *analytical*.

Most statistical studies may be thought of as having one of two general purposes. In the first place, a study may be designed to pro-

vide a basis for action on a situation as it exists. This type of study is called an enumerative study.

On the other hand, a study may be designed to provide a basis for action on the factors or on the process giving rise to a situation. This type of study is called an analytical study.

Thus, the key to determining whether a study is enumerative or analytical is an answer to the question: "Are we going to take action on the situation as it exists or on the process which gave rise to the situation?"

Example 2.12 Ordinarily wool is imported in highly compressed bales. An importer, receiving a shipment of these bales, must classify and sell the wool as first grade, or second grade, etc. Thus, for each bale he faces a decision problem.

The grade of a bale of wool must be determined by inspection. For reasons of economy and expeditiousness, however, the importer makes his decision for each bale of wool on the basis of samples of "cores" of wool rather than on the basis of a complete inspection.

Most pertinent at this point in our studies is the thought that the importer uses a sample as a basis for acting on a situation (a bale of wool) as it exists. No action is being taken on the process or causal factors in the decision problem as presented. Thus, the problem presents an enumerative study.

An enumerative study is designed to provide a basis for action on an existing situation. It deals with a fixed and finite set of observations—in other words, with a finite population.

An analytical study, however, deals with a process, and any existing information must be regarded as a sample from that process. The population for such a study would include all observations that could be made on the process if it continued to operate indefinitely under the same conditions. We may consider this set of observations as infinite. Thus, an analytical study deals with an infinite population.

Example 2.13 An industrial engineer must estimate the average time a particular polishing operation takes. This information will be used, in turn, for determining piecework wage rates.

Any statistical study of the time taken to perform this operation would deal with a process. Hence, the study would be analytical.

Theoretically the polishing operation could be repeated without limit. Thus, the number of performance-times which could be observed by the industrial engineer also is infinite. This infinite set of performance-times is the population—an infinite population.

You might note that when a study is analytical and a population is infinite, a complete count is impossible. Such a population can

never be studied completely. But this should not disturb you. It simply means that in analytical studies we *must* settle for sample data. However, even in enumerative studies we may prefer to settle for sample data, not because complete data are unobtainable but because sample data are usually more economical, all things considered. Sampling is absolutely necessary only in analytical studies, but very often it is wise in enumerative studies as well.

Whether a study is enumerative or analytical and, hence, whether a population should be regarded as finite or infinite depends on the action to be taken in the decision problem. Is action to be taken on the existing situation? Or on the factors giving rise to the situation? This fact may be further emphasized by the following illustration which shows that the same statistical information may be used for an enumerative or for an analytical study, depending upon one's decision problem.

Example 2.14 The machine department of a manufacturing company produces a certain part which is then sent, in lots of 1,000, to the company's assembly department where it is fitted to a complex piece of machinery.

Suppose a sample of parts is taken from a certain lot of 1,000 parts. Is the study enumerative or analytical? This depends on the purpose of the study—on the problem for which the sample is to be used as a basis for decision.

The purpose of sampling from a lot may be to decide whether to scrap the lot, or to inspect the lot fully and weed out all defectives, or to send the lot into production, or to do something else with *that particular lot*. In such a case the statistical study would be enumerative. Action is to be taken on a situation as it exists. The population is finite, being represented by the 1,000 parts in the lot being dealt with.

Alternately the purpose of sampling from a lot may be to decide whether or not to adjust the machine department's production process. In that case the study would be analytical. In fact, you may note that even all of the parts in a given lot present only a sample from the process. Why?

In Chapter 1 we distinguished between testing and estimation problems as one way of describing statistical studies. In this section we have classified statistical problems in another way—as enumerative or analytical. These two classifications are independent. A study may be enumerative and involve testing or estimation or it may be analytical and involve testing or estimation. Don't confuse the two classifications.

One last comment is in order. Whether it is enumerative or ana-

lytical, the purpose of a statistical study is to provide a rational basis for action in some decision problem. In both types of studies, statistics is a means to rational decisions in the face of uncertainty.

2.5 Multiobservational Problems

Thus far we have concentrated on discussing and illustrating types of statistical studies in which only one characteristic is to be observed. Such studies are important. They arise frequently in practice. Then, too, they present relatively simple situations. For these reasons, more often than not we will deal with single-observational problems. Other practical problems, however, are multiobservational. That is, two or more characteristics of a single elementary unit are observed. We shall also have occasion to refer to some of these problems.

Multiobservational problems arise in two types of situations. First of all, some statistical studies are not designed to afford a decision on only one problem. Rather, in the interests of economy as well as other considerations, they are designed to provide bases for decisions on a whole set of problems. This is typically the case in marketing and advertising research. In such studies, several characteristics may be of interest since there are several problems. Several characteristics necessitate several different observations on any one elementary unit. Hence, there are several populations of interest in the study.

Example 2.15 A real estate corporation is planning to construct several new apartment buildings. Several decision problems arise in this endeavor. Should they install electric or gas ranges? Should the kitchen floor covering be linoleum or resilient floor tile? Should the kitchen cabinets be wood or metal?

The company officials plan to do a study to provide a basis for these and other decisions that must be made. They will ask tenants in present company-owned buildings to state their preferences on ranges, kitchen floor coverings, kitchen cabinets, etc.

Thus, each tenant will be asked several questions. There are for each elementary unit—a present tenant—several observations to be made. These will provide bases for several decisions. The study presents a multiobservation problem.

Multiobservational problems also arise because occasionally, even though you are dealing with only one problem, your study will provide more information if you observe more than one characteristic. Several bits of information are, in this case, all pertinent to the one decision problem.

Example 2.16 A personnel administrator decides whether or not to pass executive trainees on the basis of several job ratings and tests. Thus, not one but several different characteristics of a trainee are observed. The problem is multiobservational.

2.6 Variates and Attributes

As has been noted, in defining a population we must specify the elementary units and indicate the characteristic (or characteristics) to be observed.

Our observations of these characteristics can be divided into two main classes: quantitative and qualitative. A *quantitative* observation is one which is expressed numerically. As the name implies, it involves some quantity. The unit of measurement may be a dollar, or an inch, or a number of employees. But quantitative observations are, by definition, always numerical. Also note that the values which the quantitative observations may assume are called *variates*.

Example 2.17 Consider the following elementary units and their characteristics, all of which may be observed quantitatively. Note also the unit of measurement and the specific numerical example of a variate.

Elementary Unit	Characteristic	Unit of Measurement	Variate
A family unit	Annual income	Dollars	\$5,340.40
A radio tube	Time to failure	Hours	953 hrs.
An employee	Monthly absences	Number	3
A preferred stock	Dividend rate	Per cent	5.5%
A ball bearing	Diameter	Inches	.06254"
A firm	Sales-inventory ratio	Per cent	752%
A machine	Age	Months	75

Quantitative observations may be further grouped as discrete or continuous. A *discrete* quantitative observation is one that can take only a limited number of values on any measuring scale. Annual income can at best be measured to the nearest cent. It cannot assume values ending in $\frac{1}{2}\phi$ or $\frac{2}{9}\phi$. The number of production workers employed by a firm can be an integral value, but it cannot be fractional. On the other hand, a *continuous* quantitative observation theoretically can assume any value on a scale. For example, measurements of length, weight, and time can be thought of as continuous. Thus, in theory, the diameter of a ball bearing might be measured to the nearest .001 inch, to the nearest .0001 inch, to the nearest .00001 inch, or as precisely as one wished.

In practice, of course, we may measure only as closely as our most precise measuring instruments will allow. As this remark implies, therefore, every continuous variate can be only approximated in the real world. It can be measured only to a finite number of places because of the limitations of our measuring instruments. However, as you will see, many times statistical theory makes the simplifying assumption of the continuity of quantitative observations. This ordinarily is an approximation which costs little in terms of accuracy.

A *qualitative* observation refers to some nonnumerical property or some attribute of an elementary unit. As the name implies, it is a quality. The terms used to describe the unit's characteristics are, in fact, referred to as *attributes*.

Example 2.18 Here are some illustrations of qualitative characteristics, a few of which we have already met in other examples:

Elementary Unit	Characteristic	Attributes
An apartment.....	Availability	Occupied or vacant
A fuse.....	Quality	Defective or not defective
A dwelling unit.....	Type of heating	Coal, gas, or oil
Flip of a coin.....	Side showing	Head or tail
An employee.....	Sex	Male or female
An employee.....	Attendance	Absent or present
An employee.....	Opinion	Very satisfied, satisfied, neutral, dissatisfied, or very dissatisfied
A firm.....	Sales level	Increasing, stable, or decreasing

In some problems, qualitative data may be regarded as numerical data, at least indirectly. For some populations we count the number of elementary units having a certain attribute. By enumerating the number of items we get a numerical result. Moreover, in such a count we score 1 for an item that has the attribute and 0 for an item that does not have it, actually cumulating the "ones" as we proceed in order to get our total. Indirectly, we are assigning the qualitative observations numerical scores, these being 0 or 1.

Now, a key point which flows from our discussion in this section is that a population may be a population of variates or a population of attributes. Before elaborating, suppose that we look at two illustrations.

Example 2.19 A small town has 750 homes. A builder, contemplating a new housing development, wishes to determine the types of heating used in the homes already built to help him decide on the heating unit he should use.

In such a problem we have the following:

Elementary unit	Home
Characteristic	Type of heating
Type of observation	Qualitative
Attributes	Oil, gas, coal, other, or none

The set of 750 attributes (one per home) is the population under consideration. If sampling were to be used, it would involve observing a sample of these attributes.

Example 2.20 A trade association has 310 members—tool and die manufacturers. It is interested in canvassing its members to determine the hourly wage rate for various job classifications, including male drill press operators, in member companies. This information is then sent to members to be used in wage negotiations.

In such a problem we have, among other things, the following:

Elementary unit	A member firm
Characteristic	Hourly wage rate of male drill- press operators
Type of observation	Quantitative
Variates	\$1.74, \$1.65, \$1.90, etc.

The complete set of variates is the population for this problem. It is a population of quantitative observations, in contrast to the one in Example 2.19 which was a population of attributes; that is, of qualitative observations.

As we have noted earlier, a statistical population must be defined clearly and unambiguously in the planning stage of a statistical study. What is to be an elementary unit and which characteristic is of interest are very important considerations. In addition, so is the matter of whether the statistical population of study is one of variates or one of attributes, and the matter of defining the attributes or the variates in clear-cut, objective, and meaningful terms. To reinforce these ideas in a specific way suppose that we consider two further illustrations.

Example 2.21 Several different types of decision problems require, for their solution, statistical studies of the *ages* of persons residing in a certain area, or of the *ages* of teachers throughout the country, or of the *ages* of employees in a certain company. For example, a firm's decision on whether or not to set up a pension plan and even its decision on how to set it up necessitate some information on the age distribution of its employees.

Thus, these statistical studies are all similar in that they involve the characteristic of age. This characteristic, however, must be more specifically defined. Is the characteristic of interest a person's age at

his last birthday, at his forthcoming birthday, or at his nearest birthday? Is the characteristic of interest age in years, age to the nearest month, or age to the nearest day? Answers to these questions depend on the purpose of a study—the uses to which the results will be put. Nevertheless, answers must be provided before we can say the population has been defined.

Since age usually is expressed numerically, studies of age usually involve a population of variates. However, suppose the decision problem only requires information on the number of persons of school age (defined as 5 to 16 years old). The study then involves a population of attributes since each elementary unit, a person, may be characterized by either of the attributes, “school age” and “not school age.”

Example 2.22 A merchandising company executive feels that through proper packaging the sales of one of his firm's products may be increased. A marketing consultant suggests a new method of packaging. The company thus has a decision problem involving two alternative courses of action: (1) continue to use the old method, and (2) introduce the new method. The company feels the use of both methods at the same time would be impractical.

A statistical study of consumer preferences is proposed. Thus, a question arises as to what type of information, if any, would help the company executive decide between the alternatives? The first thought is to define a population of attributes including only the possibilities “prefer new to old” and “prefer old to new.”

However, before any research is done, the executive should ask himself: “If I knew the number of consumers who prefer our new package to the old and the number who prefer the old to the new, could I make the required decision?” A little thought should convince him that he could not. The population has not been defined adequately nor correctly. It would be preferable if the following attributes were defined as those to be included in the population of study:

- a) Would buy new but not old.
- b) Would buy old but not new.
- c) Would buy either old or new.
- d) Would not buy either.

The complete set of these attributes for all elementary units—consumers (this too would have to be defined more carefully)—would be a far more reasonable population of study for almost obvious reasons.

We should note briefly that it is customary to introduce some special terminology for observations in multiobservational problems. When an observation is concerned with two characteristics, it is called a *bivariate* observation and the population of the observations is called a *bivariate* population. In contrast, an observation

which refers only to one characteristic is termed *univariate*. For more than two different characteristics, the terms “multivariate observation” and “multivariate population” are used. However, the terms “univariate,” “bivariate,” and “multivariate” should not mislead you. One or all of the characteristics in such a case may be attributes as well as variates.

2.7 Frequency Distributions

Thus far we have been discussing and thinking about statistical populations in terms of a complete group of observations—either variates or attributes. In a given population of study the variates or the attributes ordinarily will not be the same but will vary from one elementary unit to another. In addition, it is important to recognize that statistical studies are not concerned with individual observations on elementary units. Statistical studies are impersonal to the extent that they are not interested in whether Jones, Smith, and/or Brown have station wagons—they are interested only in how many of those elementary units have them. In short, statistical studies are designed to provide information about populations—not about specific individuals.

Generally speaking, the most detailed summary of a complete set of observations with which we ever would be concerned in a statistical study is a frequency distribution. In a broad sense the term “frequency distribution” may be applied to any forms of the data that arise when the occurrences of a particular attribute or of a succession of variate values, in a population, are counted or enumerated. Thus, a frequency distribution shows all different variates or attributes in a population and their frequency of occurrence.

For example, a census of the kinds of heating equipment in use would involve counting the number of homes that have each kind. The result would be a population of attributes described in terms of their distribution according to types of heating installations.

Such distributions of attributes are very common. Leaf through the pages of a compilation such as the *Statistical Abstract of the United States* and you will find examples by scores and hundreds. Some of these counts are made according to geographic areas, or for periods of time; others, for example, are by classes of merchandise, by occupations of persons gainfully employed, by nationality, or by sex—as well as a great many more.

Example 2.23 Many investors in stocks and bonds base their financial decisions on shifts and trends in the over-all market as well

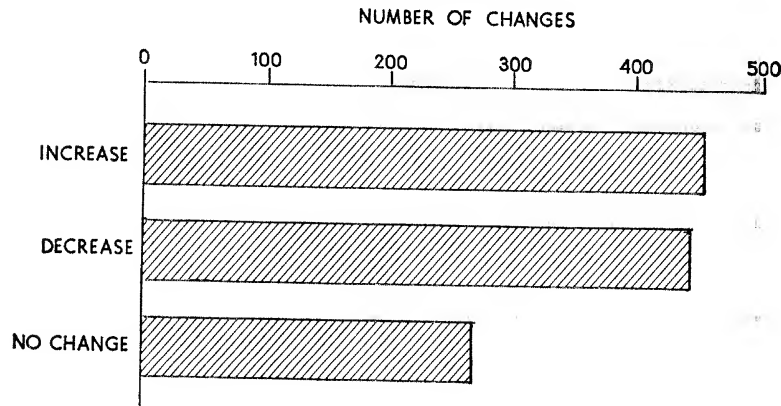
as on information about individual securities. In this connection, one population of interest is the directions of change, for a given period of time, of stocks listed on the New York Stock Exchange. In this case, an elementary unit is a stock traded on that exchange, and the population is one of attributes. The population of changes for July 1, 1958, may be shown in frequency distribution form as follows:

<u>Attribute</u>	<u>Frequency</u>
Increase.....	460
No change.....	269
Decrease.....	447
Total.....	1,176

It should be noted that this distribution presents a population for the data covered if, and only if, the data are to be used in an enumerative decision problem. If the study were analytical, dealing with the market process, the data would be sample data; they would represent a sample of but one day. Note also that the stocks included are only those traded on the New York Stock Exchange. There is always the question of whether or not this is a population adequate for a problem dealing with stocks in general.

A frequency distribution of a population of attributes may be pictured graphically in terms of a *bar chart*. The frequency distribu-

Figure 2.1
DIRECTIONS OF CHANGE FOR PRICES OF 1,176 STOCKS
NEW YORK STOCK EXCHANGE: JULY 1, 1958



tion shown in the preceding illustration is presented in that form in Figure 2.1. Note the structural resemblance between the numerical presentation and the chart of the data.

A somewhat different situation—a frequency distribution of a population of variates—is suggested if we consider a study of hourly

wage rates in a given industry by a trade association. Suppose that the information is to be used by member firms in setting wage schedules. From the reports of the member firms the association might compile long lists of wage rates. However, a moment's reflection will indicate that in that form the data scarcely will yield the information that the firms need. The interest is not in individual rates paid by individual firms but rather in the distribution of different rates in the industry as a whole. How much difference is there between the lowest rate paid and the highest? Do many of the rates tend to cluster around some norm? Is the tendency very marked, or only moderate? Is the norm at a high or a low level in the scale of rates?

If the wage rate data are not arranged in some special order, or given a definite structure, the very mass of the numerical facts will tend to obscure whatever significant information they may contain. Useful answers to such questions as those above will be difficult to obtain.

A simple but effective procedure may be employed to reveal the internal structure of the data. It is carried out as follows: (1) set up quantitative groupings of wage rates that will serve to break up the whole range of rates into a limited number of classes; and (2) count the number of rates that fall within the limits of each class. Typical results are illustrated in the next example which covers a situation analagous to the one we have been discussing.

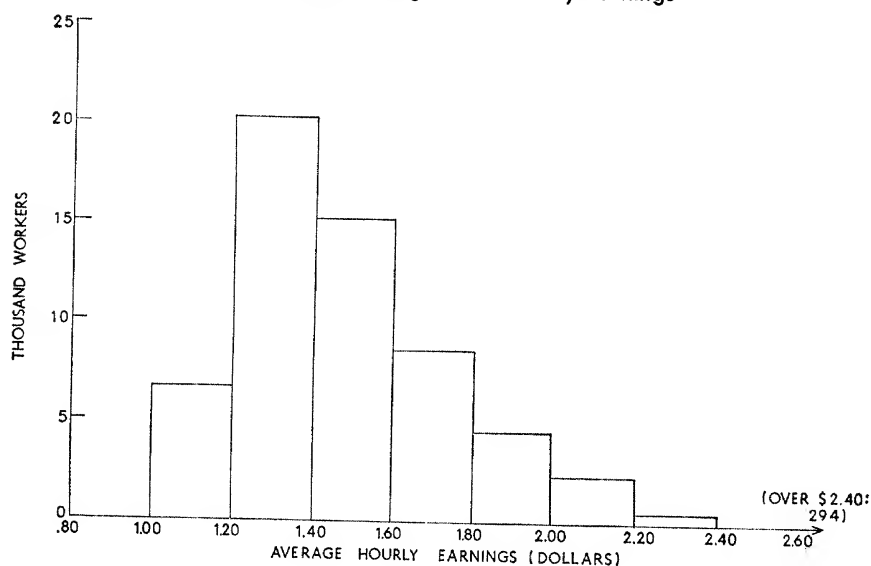
Example 2.24 The *Monthly Labor Review*, May, 1958, presents a study of average straight-time hourly earnings for production workers in woolen mills as of September, 1957. Data are given for the United States as a whole and for various geographic regions. To serve the purposes of this example only the distribution of workers in the New England region has been used, together with that for the United States. The figures shown below have been derived by computation from those that appear in the original source:

Average Hourly Earnings (in Dollars)	Number of Workers	
	United States	New England
1.00 and under 1.20	6,537	2,009
1.20 " " 1.40	20,258	9,128
1.40 " " 1.60	15,135	9,453
1.60 " " 1.80	8,598	4,815
1.80 " " 2.00	4,711	2,511
2.00 " " 2.20	2,532	1,329
2.20 " " 2.40	648	266
2.40 or more	294	59
Total	58,713	29,570

Note that as a result of dividing up all of the wage rates among the various classes we have a *frequency distribution*. In place of a large number of individual rates we have a compact description of a population of variates in terms of the frequency of their occurrence and of the essential characteristics of their distribution. In addition, although it ignores which elementary units have which characteristics, it summarizes the information of statistical interest—how many elementary units have certain characteristics. At a later stage of our study we shall give somewhat more attention to the techniques applicable to the construction of frequency distributions.

A frequency distribution of a population of variates may be shown graphically in the form of a histogram (or column diagram). For example, the frequency distribution for the United States shown in the

Figure 2.2
DISTRIBUTION OF PRODUCTION WORKERS IN WOOLEN MILLS
UNITED STATES: SEPTEMBER, 1957
By Average Straight-Time Hourly Earnings



preceding illustration is presented as a histogram in Figure 2.2. It should be noted that a direct relationship exists between the areas of the diagram and the frequency, or number, of occurrences. The area of the whole figure is proportionate to the total number of observations, and the area of each column individually is in proportion to the frequency of the class it represents. Thus the histogram very effectively pictures the form and character of a population's distribution.

At times we may prefer to convert actual, or absolute, frequencies to a percentage form. This results in a *relative frequency distribution*

in which the frequency of each class is expressed as a ratio to the total number of cases. One reason for doing this arises in a situation in which we want to make a comparison between two population frequency distributions. In such a case, it may be misleading, or at least difficult, to compare actual frequencies because one population includes far more observations than the other. Relative frequencies may indicate more clearly the differences and/or similarities in the two populations of interest.

In Example 2.24, if we attempt to compare the distribution for New England with the one for the United States generally, we find so much numerical difference in the respective frequencies for given classes of hourly earnings that the result is without any significance. We almost inevitably must ask ourselves such questions as: "Are 8,598 workers (at the \$1.60 to \$1.80 level) a greater or a lesser proportion of all United States wool production workers than 4,815 are of those in New England?" In other words, we must think in relative rather than in absolute terms. The next example should help to make clear the greater effectiveness of percentage distributions in situations of this sort.

Example 2.25 Following are relative frequency distributions of production workers in woolen mills by average straight-time hourly earnings. The basic data are the actual distributions of Example 2.24. (Items do not add to exactly 100 per cent due to rounding.)

Average Hourly Earnings (in Dollars)	Per Cent of Total Workers	
	United States	New England
1.00 and under 1.20	11.1	6.8
1.20 " " 1.40	34.4	30.9
1.40 " " 1.60	25.7	32.0
1.60 " " 1.80	14.6	16.3
1.80 " " 2.00	8.0	8.5
2.00 " " 2.20	4.3	4.5
2.20 " " 2.40	1.1	.9
2.40 or more	.5	.2
Total	100.0	100.0

In this section we have viewed the frequency distribution as a device mainly for summarizing a statistical population. It must be stressed again, however, that in most decision problems we cannot construct frequency distributions of populations—simply because complete information rarely is available for such a task. Nevertheless, the frequency distribution is an important concept. Occasionally

a population is available and may be so summarized (as in this section's illustrations). Also, you now should realize that in many problems, even though only a sample of observations is to be obtained, thinking about the population is necessary in order to focus on the type of data required to make a decision. In addition, thinking about the population in terms of a frequency distribution points up in which form these data are needed in many problems. Then, too, frequency distributions are useful in other connections in statistics—for example, to summarize sample data for further analysis. In our study of statistics we will meet several different situations in which we may use the concept of a frequency distribution.

2.8 Population Models

Frequency distributions of variate populations, as you may well imagine, can assume an almost endless variety of shapes and sizes. Nevertheless, it is possible to classify them into a few general types. For example, in Figures 2.3–2.8 we have sketched a series of histograms to illustrate some of these general types of frequency distributions of populations which one meets in statistical practice.

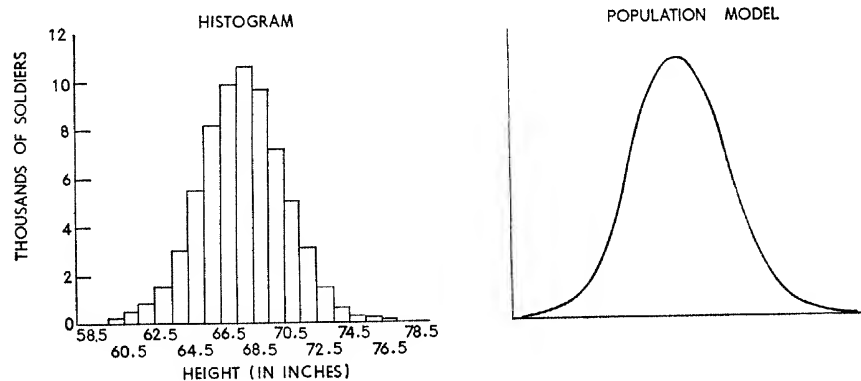
Accompanying each histogram is a smooth continuous curve which depicts the more salient features of the frequency distribution. We will refer to these smooth curves as *population models*. In using smooth curves such as those drawn, a statistician resembles a caricaturist who by a few bold strokes of his pen can so capture the essence of an individual that he is easily recognizable. Although such smooth distributions are never observed in practice, they present a simplified method of describing populations. The term “population model” is, in fact, suggestive. It connotes that the smooth curves are models rather than perfect replicas of populations. These models or curves can be represented by mathematical equations, and as we shall see later in the text, the use of such equations offers a distinct advantage in the development and application of statistical theory.

We may turn now to a discussion of the illustrations of the populations and their models provided by Figures 2.3 through 2.8. Figure 2.3 presents the frequency distribution of the heights of soldiers inducted into the Army in 1943. This distribution is low at the two extremes but rises fairly steadily to a single peak which is near the center of the distribution. There are very few extremely short soldiers, very few extremely tall ones, and a larger number of soldiers of moderate height. The most typical height falls between 67.5 and 68.5 inches. The pattern of variation to the left and to the right of

this most typical height is almost *symmetrical*. In the diagram accompanying the histogram we have drawn the smooth curve (population model) which describes these main features of the distribution.

The model, called the *normal curve* or the *normal distribution*, is of special interest to statisticians for many reasons, and we shall discuss it in great detail in later chapters. It satisfactorily describes many populations for a large variety of physical characteristics and

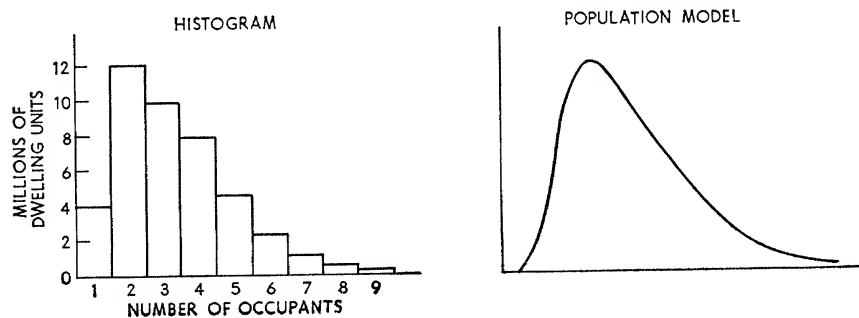
Figure 2.3



for several measures of educational and mental attainment such as IQ tests. In addition, many quality characteristics—diameters, weights, lengths, etc.—of manufactured product also will be found, in many applications, to be normally distributed, or at least approximately so. Several other statistical populations involving business and economic data also are approximated closely by the normal distribution.

Figure 2.4 involves a frequency distribution of the number of occupants living in dwelling units in the United States. This distribution has a single peak, as the last population did, but it is located at

Figure 2.4



the lower end of the distribution. In other words, the branches of the curve are not symmetrical about the peak. Such *asymmetrical* distributions are known as *skewed* distributions. Furthermore, since the longer tail of the distribution is toward the right, we say that it is *skewed to the right* or *positively skewed*. The accompanying population model which summarizes these pertinent facts is also commonly encountered when one deals with business and economic data. It might be used, for example, to depict populations of personal incomes, corporate incomes, sales, commodity prices, assets of companies, etc.

Figure 2.5

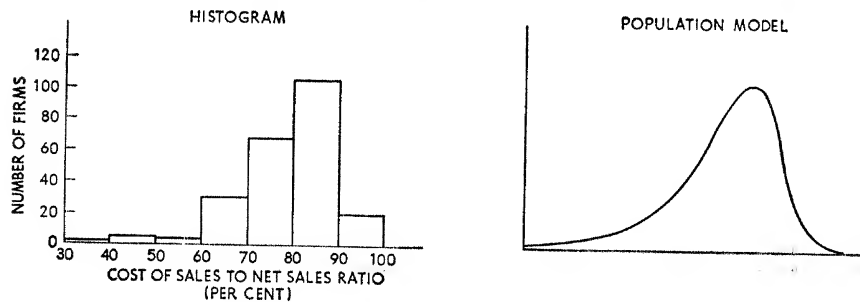
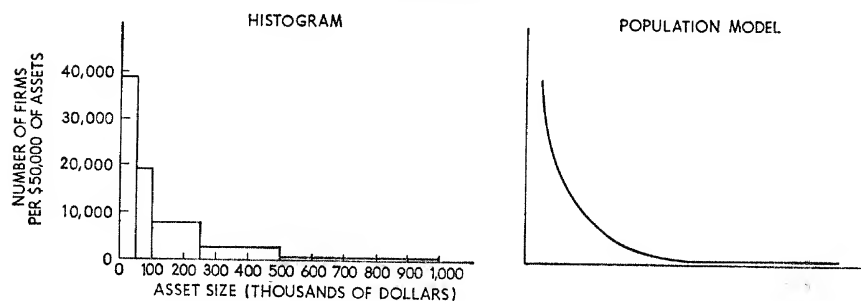


Figure 2.5 represents the distribution of the ratio of cost of sales to net sales for a segment of the plastics industry and the corresponding population model. Note that this type of population is also skewed, but since its longer tail is to the left, we say it is *skewed to the left*, or *negatively skewed*. Populations of variates which have an upper bound may, as a general rule, be approximated by this model. Thus, the upper bound of the ratio of cost of sales to sales is roughly 100 per cent because firms could not stay in business for any length of time if their cost of sales equaled or exceeded their sales.

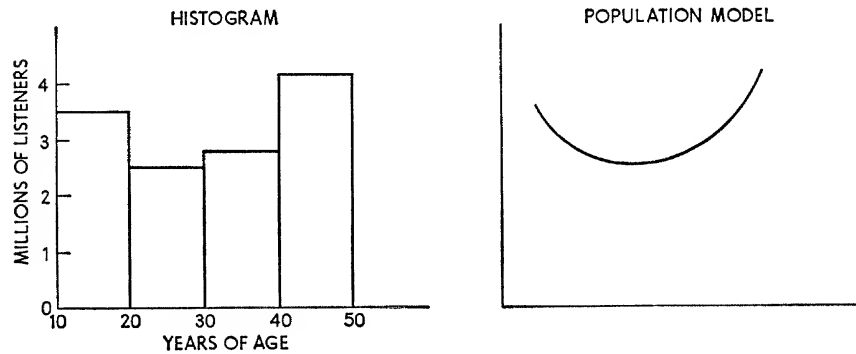
The distribution of the number of corporations filing corporate in-

Figure 2.6



come tax returns classified by size of their assets is shown in Figure 2.6 together with a population model representing it. Because of its shape such a model often is referred to as a (reversed or inverted) J-shaped distribution. This model also is descriptive of various types of "failure" data. For example, a population of times to failure of retail businesses would have the same general shape as the model in Figure 2.6. Most firms, if they fail, fail rather quickly. The longer a firm is in business, the less its chance of failing. Hence fewer firms fail after having been in business several years; still less fail after having existed for many years.

Figure 2.7



There are several other less common population models of some interest in statistics. For example, Figure 2.7 portrays a U-shaped frequency distribution of ages of viewers of a Jack Benny program and its appropriate population model. This type of distribution is far from common in practice, but obviously it does occur.

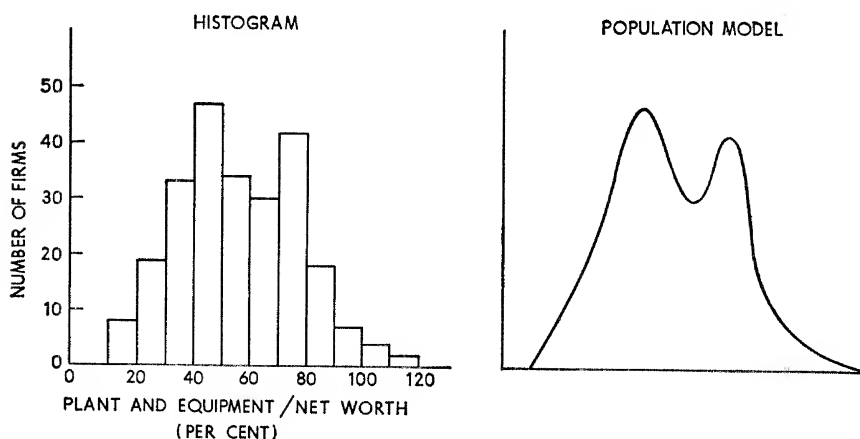
Some distributions have only one peak and are called unimodal. On occasion we may encounter distributions with two (or more) prominent peaks—bimodal (or multimodal) populations. Thus Figure 2.8 portrays the distribution of the ratio of the value of plant and equipment to net worth for a group of plastic molders in the United States, and a model which describes this series. The two peaks occur because the distribution combines ratios for small and large firms in the industry. While the ratios of the larger firms are concentrated between 40 and 50 per cent, the ratios for smaller firms are concentrated between 70 and 80 per cent.

Another example of a bimodal population would be a frequency distribution of the heights of males and females. Characteristics of the work produced on different shifts but presented in one distribution might also assume this form. Thus, one reason for bimodal pop-

ulations is that they include two sets of somewhat different characteristics (e.g., male heights and female heights).

The various population models we have been discussing are, of course, quite useful in statistical work. Why? First of all, they offer a simple method of describing a statistical population. Thus, if we say a population is very skewed to the right, a general picture of the population emerges. Similarly, the use of the terms "symmetrical,"

Figure 2.8



"unimodal," "U-shaped," or "J-shaped" call to mind essential features of frequency distributions of populations.

Second, a population model is very often an important factor in determining the pertinent fact or facts we have to know about the population in order to reach a decision on a course of action in a given problem. We will discuss a few uses of models in this connection in Chapter 3.

Third, we have already stated that a key problem in statistics is to keep the cost of obtaining observations within economic limits—that is, as low as economically feasible. Population models are important in this connection because the design of statistical studies is often simplified by some knowledge of a population's shape. We can see now that such information is often available on the basis of a priori reasoning. For example, even if we had never seen the results of a census of retail trade, we could deduce that the distribution of sales volume would be very skewed. It is common knowledge that there are many small stores with a relatively small sales volume but few very large retail outlets whose sales run into hundreds of millions of dollars.

In conclusion, it may be noted that although we have limited our discussion in this section to *population* models, many of these same smoothed curves are used to describe other types of frequency distributions.

2.9 You Should Now Know That

The first step in the planning stage of a statistical study, or at least the first thing to do after identifying the problem, is to define the population.

A population is the totality of all pertinent observations that might be made in a given problem.

Defining the population is an objective and clear-cut way of stating what type of information is needed to make a decision.

Statistical studies involve observations of characteristics of elementary units.

The choice of an appropriate elementary unit and of the characteristic of interest are two important steps in defining a population.

In many statistical problems a frame is necessary to identify the elementary units to be studied. A frame is a list or map or some other concrete way of identifying those units, and it provides a means of access to the population.

A frame is as essential for a complete census as for a sample study.

In practice, frames often are subject to various imperfections.

Populations may be finite or infinite, depending on whether the statistical study is enumerative or analytical; this, in turn, depends on the purpose of the study.

Some statistical problems are multiobservational since they deal with observations on more than one characteristic of an elementary unit.

Observations may be quantitative or qualitative. If quantitative, the values the observations may assume are called variates; if qualitative, the properties observed are called attributes.

A frequency distribution shows the different observations in a population and their frequency of occurrence.

A population model of a variate distribution is a smoothed curve representing the salient features of the distribution.

2.10 You Should Now Be Able to Solve

1. Comment on the following statement: "The determination of the statistical population to be studied is not wholly a statistical problem."

2. For each of the following proposed studies indicate (i) an elementary unit, (ii) a characteristic of interest, and (iii) whether observations would be quantitative or qualitative. Explain each answer with reference to the specific problem.

- a) A college is planning a survey to determine the major fields of all matriculated students; the information is to be used in deciding on future course offerings.
- b) A municipal housing authority is planning a study to determine the demand, among low-income families, for 1-, 2-, 3-, and 4-bedroom apartments—the study to provide a basis for planning its new construction.
- c) The U.S. Bureau of the Census is planning its Census of Manufactures which is to include data on the total number of production workers employed in manufacturing establishments.

3. Discuss the suitability of the frame suggested in each of the following situations:

- a) The telephone directory as a frame for a study by a city councilman to determine the attitudes of voters in a given city towards a proposed bond issue.
- b) Instructors' class rolls as a frame for a study to determine the number of absences from classes by freshmen, sophomores, juniors, and seniors at a university. The study is to be used as a basis for a decision on the university's policy on the allowed number of absences.
- c) A list of past purchasers of EZ-shave electric razors (obtained from warranty registration cards returned to the company) as a frame for study on consumer preferences for a new model razor.

4. A manufacturer of electric switches suddenly experiences a sharp increase in the number of defective wall switches returned by customers. He plans to check his manufacturing operations as a basis for deciding on whether or not to make any adjustments in the operations. Is the proposed study enumerative or analytical? Explain.

5. *American Standard, Z1.3-1958*¹ includes the following comments:

The manufacturer's inspection of his own product serves two purposes:
Purpose A. To provide a basis for action with regard to the product already at hand; as for instance, to decide whether the particular article or lot of product at hand should be allowed to go out, or some alternative disposition made (inspected further, sorted, repaired, reworked, scrapped, etc.).

Purpose B. To provide a basis for action with regard to the production process, with a view to future product; as for instance, to decide whether

¹ *American Standard, Control Chart Method of Controlling Quality During Production, Z1.3-1958* (New York: American Standards Association, Inc., 1958), p. 5.

the process should be left alone, or action taken to find and eliminate disturbing causes.

Which of these situations presents a finite population of study? Which an infinite population of study? Explain.

6. A magazine plans to conduct a survey to provide information about its readers that will aid advertisers in reaching decisions on the preparation of advertising copy. Why might such a study be multi-observational?

7. A statistician once stated: "I have never yet seen an inspection problem which would not benefit from the point of view that the product to be inspected was a frequency distribution."² What do you think prompted this comment?

8. What type of population model would best represent the following populations? Describe as symmetrical, or U-shaped, or skewed to the left, etc. Sketch the smooth curve you have selected, and, in each case, explain the reasons for your choice.

- a) The monthly rents of 3-room apartments in Chicago.
- b) The diameters of 2-inch dial knobs produced by a plastics firm.
- c) Scores on College Entrance Board examinations.
- d) Ages of unemployed persons 14 years old and over.
- e) The ages of readers of a science-fiction comic book.

2.11 You Will Also Find That

Many of the basic concepts regarding statistical populations, elementary units, characteristics, and observations are discussed and illustrated in a number of statistical textbooks including the following:

SPROWLS, R. CLAY. *Elementary Statistics*, pp. 8-10. New York: McGraw-Hill Book Co., Inc., 1955.

MCCARTHY, PHILLIP J. *Introduction to Statistical Reasoning*, pp. 8-20. New York: McGraw-Hill Book Co., Inc., 1957.

For an extensive discussion of the use and the imperfections of frames in several different types of statistical studies, see:

YATES, FRANK. *Sampling Methods for Censuses and Surveys*, pp. 48-49, 60-87. New York: Hafner Publishing Co., 1949.

The terms "enumerative" and "analytical" are due to Professor W. Edwards Deming. For further illustrations see:

² G. D. Edwards, cited in Eugene L. Grant, *Statistical Quality Control* (2d ed.; New York: McGraw-Hill Book Co., Inc., 1952), p. 407.

DEMING, W. EDWARDS. "On the Distinction between Enumerative and Analytical Surveys," *Journal of the American Statistical Association*, Vol. XLVIII (1953), pp. 244-47.

DEMING, W. EDWARDS. *Some Theory of Sampling*, pp. 247-50. New York: John Wiley & Sons, Inc., 1950.

The concepts are also discussed in terms of "existent and hypothetical populations" in:

YULE, G. UDNY, and KENDALL, M. G. *An Introduction to the Theory of Statistics*, pp. 332-33. 13th ed. London: Charles Griffin & Co., Ltd., 1948.

Almost every statistics textbook discusses frequency distributions in more or less detail. Two good sources on this topic are:

NETER, JOHN, and WASSERMAN, WILLIAM. *Fundamental Statistics for Business and Economics*, pp. 174-87. New York: Allyn & Bacon, Inc., 1956.

MILLS, FREDERICK C. *Statistical Methods*, pp. 40-64, 74-85. 3d ed. New York: Henry Holt & Co., Inc., 1955.

CHAPTER · 3

Decision-Parameters

3.1 Introduction

Chapter 1 dealt with some of the basic principles of decision making. It also included the thought that some decision problems are statistical. It was pointed out that statistical decision problems are characterized by a decision maker's uncertainty about some pertinent conditions surrounding his problem situation; that is, by his uncertainty about the true state of the world. They also are characterized by the fact that observations or numerical information will help reduce this uncertainty.

Chapter 2 focused on one step in statistically defining a statistical decision problem—on the question of how one goes about rigorously specifying the type of information needed in order to reduce or eliminate that uncertainty. This step was called defining the statistical population.

A second step, the topic of this chapter, is called specifying the decision-parameter and presents a way of statistically describing the condition of one's problem situation. As you will see, specifying the decision-parameter is intimately tied to the definition of a population. In practice, they are not two steps but one.

First note, however, that even though a decision maker knew his whole population of interest, a wise decision would not flow from this mass of observations—for these data, in themselves, would not tell him what he needs to know in comprehensible form. Usually, to take action, he would like to know some summary property of the population such as the population *average* or the population *proportion* (two things which we will study in detail below). For example, even if the population of your statistics quiz grades were known, your final mark would be based on some average of these grades. Similarly,

even if a personnel administrator knew how each one of his employees felt about a social activities program, he ordinarily would base his decision on the proportion of employees who favored (or who did not favor) the program.

The average and the proportion in these situations represent ways of describing conditions pertinent to the decision problems. Hence, they represent ways of statistically describing states of the world. Such measures may be called decision-parameters.

Thus, a *decision-parameter* is a property of a population which is needed as a basis for making a decision. It summarizes, in a single number, information regarding the population. It presents, for many decision problems, a method of describing states of the world. In this chapter we will study, in detail, several important decision-parameters, including those already noted. To further reinforce this general idea, however, let's first look at two illustrations of decision-parameters and their role in decision problems.

Example 3.1 Every college receives many applications from candidates for admission. A college thus faces a decision problem on whom it should admit. One population of interest to admissions directors involves scores on the College Entrance Board examinations. However, the whole population of scores is not necessary in order to decide who should be admitted; only certain properties of the population are needed. Often the property of the population needed is the *90th centile* (or 90th percentile)—that score under which 90 per cent of the scores fall and over which 10 per cent of the scores fall. Thus, students with scores over the value of this decision-parameter are admitted; others rejected.

Other schools, of course, may use the 75th centile or some other centile as their decision-parameter. Whatever their choice, however, note that it describes states of the world. If the centile is known, it provides a college with knowledge of the true state of the world and, hence, a basis for deciding who should be admitted.

Example 3.2 A baby food products company has a mailing list of customers. The list is used to identify customers to whom coupons redeemable for samples or sales promotional literature are forwarded. Names, derived from many sources, are continuously added to the list. Because of this variety of sources, the number of duplicate, triplicate, and quadruplicate names builds up in the list, and from time to time the company sorts out these repeats. This, however, is a somewhat costly and tedious operation.

At any time the company faces a decision problem: to sort or not to sort the list. The population consists of the names in the list. A frequency distribution of the population, then, might show the number of names occurring once in the list, occurring twice, occurring

three times, etc. However, this would be more detailed information about the population than is needed to make the decision on whether or not to sort out the list. What is needed is the number of different names in the list.

Thus, suppose they have a list of 250,000 (this is assumed to be known). They would not need to know that 180,000 names appeared but once, 20,000 appeared twice (representing 40,000 entries on the list), and 10,000 appeared three times (representing 30,000 entries on the list). They would be interested only in the fact that 210,000 of the 250,000 names were unique, while 40,000 were elsewhere in their list.

Thus, the total number of different names in the list is the correct way of describing the problem situation. It represents states of the world and is, hence, an example of a decision-parameter.

Thus, as these two illustrations indicate, the nature of the problem determines the appropriate decision-parameter. Sometimes a total number is needed, sometimes a proportion, sometimes an average, and sometimes some other property of a population. Sometimes, of course, management might wish it knew two or more properties of the population before making a decision on a problem. For example, a manufacturer of light bulbs may wish to know the *average* quality and the *variability* of his product, in which case the problem involves two decision-parameters.

Any statistical population usually has many properties—an average, a minimum value, a maximum value, various proportions, etc. Every such property is called a *parameter*. Not every parameter is, however, a *decision-parameter*. One never need know every property (parameter) of his population in order to make a decision—just certain important properties, namely, those decision-parameters of interest. Thus, the key question in determining the appropriate decision-parameter(s) is always: “What do we have to know about the population in order to make a decision?”

In some situations the answer to this question will point up the fact that we need a frequency distribution of the population. It may be noted, however, that a frequency distribution is a set of decision-parameters since each frequency is a property of a population. This ordinarily is the most detail about a population one would need in order to reach a decision.

Example 3.3 A haberdasher must decide on the proportion of his pajama inventory that should be in sizes A, B, C, D, and E. The pertinent fact for his decision problem is the relative frequency distribution of his male customer's pajama sizes. Thus, after having de-

cided on how much of an inventory to carry (this would depend on other information) he might, given the distribution of sizes, buy pajamas in proportion to his population's relative frequencies. For example, if he were able to determine that 10 per cent of his customers wore size A, he might buy 10 per cent of his stock in that size.

Knowledge of an average size or the most commonly requested size rarely would be helpful to a small haberdasher. He must know the entire distribution of sizes.

Example 3.4 A company considers setting up a group life insurance program for employees. It contacts an insurance agent and requests information on the initial cost of such a program. This presents the agent with an estimation problem. In order to determine the premium, he needs to know the frequency distribution of the ages of employees (or the distribution of birthdates). He could then apply rate tables to determine the premium.

Note that knowledge of the average age of employees would not be sufficient information for the agent to decide on the premium. For example, suppose the agent knew the average age of 2 employees was 50. The premium would be different if their ages were 25 and 75 rather than 49 and 51. Life insurance rates are not proportional to ages.

Thus, in some problems the population frequency distribution is the fact about a population needed by a decision maker. This, then, is the way we describe states of the world. Each state of the world is some frequency distribution; the *true* state of the world is *the true*, but often unknown, frequency distribution.

Of course, the idea of a decision-parameter does not eliminate one basic problem—the fact that often complete information about a population is unavailable and, therefore, so are those pertinent facts called decision-parameters. Once again, however, it is necessary to realize that we must first determine what information we need—and do this before worrying about how we are to obtain it. Statistical methods and theory provide a basis for getting some information via sampling, but one must first know for what to sample and why he is sampling. Thus, specifying the decision-parameter is an important step to take in the planning of a statistical study; it should be taken coincidentally with the definition of the population.

3.2 The Proportion

It may surprise you to find that one of the most useful of all decision-parameters is also one of the most simple: the *proportion*.

To refresh your recollection on the nature of proportions consider a housing study of occupied and unoccupied apartments. If the pop-

ulation were known, it might be found that of 1,250 apartments in a specified area, 14 were unoccupied. Hence, the proportion of unoccupied apartments would be

$$\frac{14}{1,250} = .0112 = 1.12\% .$$

Alternately, if 14 apartments are unoccupied, 1,236 must be occupied, and the proportion of occupied apartments would be

$$\frac{1,236}{1,250} = .9888 = 98.88\% .$$

It will be convenient, in our study of the field of statistics, to adopt various symbols to represent different quantities or measures. You may think of these symbols as names or synonyms. Among the different quantities we will want to name symbolically are decision-parameters, and to keep our name calling reasonably consistent, we will adopt Greek letters for almost all decision-parameters.

As a start, suppose we agree that the population proportion is to be known as π (the Greek letter *pi*). In any specific problem, π is whatever proportion we define it to be. Thus, in the illustration noted two paragraphs ago, if we define π as the proportion of unoccupied apartments, $\pi = .0112$; if we define π as the proportion of occupied apartments, $\pi = .9888$.

Note that in this section we are studying the proportion as a *property of a population*. Moreover, we are studying the proportion as a decision-parameter—that is, as a property of the population which would serve, if known, as a basis for decision. Of course, in a given problem, π is the decision-parameter of interest only if that property of the population is needed.

Despite its simplicity, the proportion arises as the decision-parameter in many problems; that is, it represents, in many situations, the key bit of information a decision maker would need in order to make a choice among alternative courses of action.

Example 3.5 A company is considering the simplification of several types of jobs. Pertinent information on these jobs in defining new work loads is the time spent on supervisory, as opposed to non-supervisory, work. Exactly which time of the day (e.g., 9:01–9:22 A.M., 9:30–9:52 A.M., etc.) is spent on supervisory work ordinarily is unimportant in this type of decision problem. Usually, the pertinent fact about the population of study is the *proportion* of time spent in supervisory work. Thus π is, for such problems, a decision-parameter.

Example 3.6 A firm has a policy of attempting to fill orders within a week's time. In the past it has determined that only about 5

per cent of the orders do not meet the specified time limit, and this is a level management feels is satisfactory. Recently, however, a few complaints about delays in shipments have been received. In view of the complaints the firm wants to determine if its order-filling operation has slowed down or has otherwise deteriorated.

The decision problem is to retain the present system or to revise it. The appropriate decision-parameter is π : the proportion of delayed orders the present system allows. It is, as the problem has been formulated, the key bit of information the firm needs in order to choose between its alternatives. For example, management may answer the question of how the information is to be used by stating that a value of π that is more than 5 per cent requires revamping of the present system. This, of course, depends on present managerial policy. In any case, π is the important property of the population for the problem.

A decision-parameter presents a way of describing states of the world. Thus, in problems in which the proportion is the key decision-parameter, it represents a description of states of the world, and the true value of π represents the true, but ordinarily unknown, state of the world. A state of the world, as you should recall, is a complete description of one's problem situation. It is important because the consequences of any decision depend not only upon the decision one makes but also on which state of the world is true. In problems in which the proportion is the key decision-parameter, then, the consequences of a decision depend on the decision made and the value of π as well.

Example 3.7 A manufacturer periodically purchases piston rings in lots of 2,000. As each lot is delivered to his plant, he must decide between two courses of action: (1) inspect the lot of product fully to weed out the defective parts, or (2) immediately send the full lot of parts on to his production line without any inspection.

Inspection costs are about 5¢ per item, or \$100 for a full lot of 2,000 parts, and we shall assume that 100 per cent inspection eliminates all defectives. Even if alternative (2), no inspection, is adopted, the defective parts will be discovered by a final test at the end of the production line, and may be replaced at that time. However, the cost per unit of finding and changing an unsatisfactory part is an additional \$2.00 per unit.

If it were known that 54 of the piston rings in a particular lot are defective, then π , the proportion of defective piston rings, would be $\frac{54}{2,000} = .027$. It follows that under these conditions, 100 per cent inspection would cost \$100; no inspection would cost \$108 ($\2.00×54 defectives). Obviously, in this case, 100 per cent inspection is the more desirable alternative.

However, under other conditions no inspection may be the preferable alternative. For example if π were equal to .01 (20 defectives out

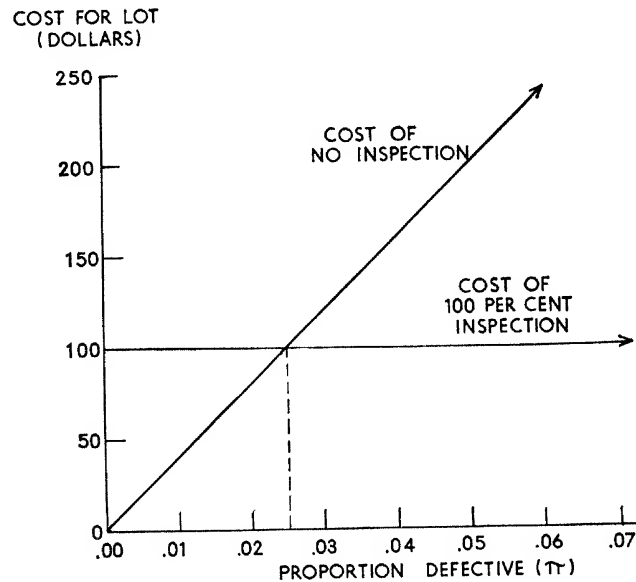
of the 2,000), then the cost of using 100 per cent inspection would still be \$100, but the cost of passing the lot uninspected would be only \$40 ($\2.00×20 defectives).

A more complete schedule incorporating these results would appear as follows:

π	Cost of 100% Inspection	Cost of No Inspection	Preferable Alternative
.010	\$100	\$ 40	No inspection
.020	100	80	No inspection
.025	100	100	Either
.030	100	120	100% inspection
.040	100	160	100% inspection
.050	100	200	100% inspection
.100	100	400	100% inspection

Graphically, these costs may be shown for various values of π , as in Figure 3.1. By inspection of this figure it can be concluded that if π for any lot were known, the manufacturer would use the following rule: If π were less than .025, he would use no inspection; but if π were greater than .025, he would use 100 per cent inspection.

Figure 3.1



The foregoing analysis in terms of costs has pointed up the fact that π is the appropriate decision-parameter. Note, in particular, that the consequences of a decision depend not only on the alternative selected but on the value of π as well. Thus, if it were available, knowledge of π would permit rational action.

Unfortunately, π rarely can be made available—to make it available would require complete inspection, in which case the decision problem would vanish. When π is unknown, can the decision be based on sample information? The answer is yes, though it must be qualified, as is explained in later chapters.

The proportion occurs as a decision-parameter in both enumerative and analytical studies. This means that π may be the pertinent fact about a finite population or about an infinite population. Its use in the case of enumerative studies and, hence, with finite populations is clear cut. However, the notion of a proportion in analytical studies and, thus, with infinite populations may trouble you. How can π be computed in such a case? For example, we never can count all the defectives a production process might produce and divide this by the total number of parts it would turn out in the long run, thereby calculating the proportion of defectives. Such data are impossible to obtain—hence, such a calculation cannot be made.

However, consider the population of all possible outcomes associated with flipping a coin. While we cannot count the total number of heads or tails in such a population, we can imagine the relative frequency of heads, after a large number of trials, to be approximately .50 (if the coin is unbiased), or perhaps .55 (if the coin is somewhat biased). Similarly, we can think of the proportion of defectives a machine produces in the long run as possibly equal to .03, or .012, or .054, the exact proportion depending on the efficiency of the operation. Intuitively, what we are considering is the mathematical notion of a limit—what proportion would occur in the long run.

The whole matter of calculating the population proportion exactly is usually academic. It is impossible to calculate π in analytical studies because complete information is impossible to obtain. In addition, π rarely can be calculated even if we are dealing with a finite population—in most of these cases complete information also is unavailable. In both types of situations, statistical methods and theory provide ways of estimating π or of otherwise acting in the absence of complete information about the exact value of that decision-parameter. But, in the planning stage of a study it is still necessary to focus on the appropriate decision-parameter and to answer the question of how it is to be used.

One other extension of the idea of a proportion as a decision-parameter involves the thought that in some problems not only one but several population proportions may represent the information

needed by a decision maker. For example, a marketing problem may require information on the proportion of consumers preferring brand A to all other brands, the proportion preferring brand B to all others, and the proportion preferring brand C to all others. Or, an economic survey of businessmen's expectations about the trend of prices may involve obtaining information on the proportion of expectations of an increase, the proportion of expectations of no change, and the proportion of expectations of a decrease. Such problems simply present situations in which there are several decision-parameters.

What is the aim of a statistical study in any problem of the type discussed in this section? Generally, it is the same aim as that of any statistical study—to provide a basis for decision. Specifically, it is to provide a basis for decision even in the absence of knowledge about the true value of π —the true value of the decision-parameter which represents the true state of the world. This it may do by providing sample information, within economic limits, which reduces the uncertainty about the true value of the unknown population proportion, thereby permitting rational action even in the face of uncertainty. Of course, if the population proportion were known, or could be made known, an appropriate decision could be made without any uncertainty about π .

3.3 The Aggregate

In this section we shall turn to another decision-parameter which occurs in a number of statistical decision problems. This decision-parameter is based simply on the idea of a total or a sum of numbers. In the field of statistics, however, this measure generally is called the *aggregate*.

Consider a population of variates; that is, a total set of numerical observations that may arise in a given decision problem. As its name implies, the aggregate is, in such a case, the sum of all the variates in the population.

Thus, in this section we shall be studying the aggregate as a property of a population. An important question in this connection is: "In what types of problems does the aggregate represent key information about a population?" That is, when is the aggregate the decision-parameter of interest in a problem?

The general answer is that the aggregate is the key decision-parameter when it adequately describes the states of the world in a decision problem—when the consequences of a decision depend on the

decision made *and* the value of the population aggregate as well. Let's reinforce these general ideas with some specific illustrations:

Example 3.8 An auditor of the books of a small textile firm must verify the company's total of accounts receivable. The firm has 20 such accounts. If the population of *true* balances for these accounts were known, they might be listed as follows: \$1,225.04, \$218.96, \$.00, \$450.00, \$1,100.50, \$654.05, \$450.00, \$876.18, \$.00, \$.00, \$98.00, \$.00, \$768.13, \$56.08, \$1,106.22, \$875.00, \$.00, \$45.55, \$2,100.00, \$67.05.

This mass of data, in itself, is not what the auditor needs to make his verification decision. By the very definition of the problem, he needs to know the aggregate of this population. Again assuming complete data on the true balances are available, this aggregate is:

$$\$1,225.04 + \$218.96 + \$.00 + \dots + \$67.05 = \$10,090.76.$$

However, since such a total would not be available in the planning stage of a study, the auditor should ask himself at that time what he might do if he knew the value of this aggregate? For example, suppose you were the auditor and the firm listed accounts receivable as \$10,090.22 on their books. If you were later to find, as above, that the aggregate was \$10,090.76, what course of action would you take? Would you certify their figure or refuse to do so? What if the book balance was \$12,897.22 and you were to find, after your study, that the true balance was \$10,090.76?

In the last section we gave the name π to the population proportion. In the same spirit, we may also call the population aggregate A (simply the first capital letter of the English alphabet, but also, if you prefer, the Greek capital letter *alpha*.) Thus, in Example 3.8, $A = \$10,090.76$.

Example 3.9 The policy of a large retail store is to stock \$15,000 worth of costume jewelry for the Christmas season. Some stock, the value of which is unknown, is on hand for this year's coming season. The present stock may be thought of as a population aggregate and symbolically represented by A .

The decision problem is to decide how much, if anything, to purchase. The aggregate is, in this situation, the key decision-parameter. It describes the problem situation; hence, it describes states of the world. The consequences of any decision depend not only on the alternative selected but on the value of A as well.

Thus, if the store determines the value of A , they obviously can make a purchasing decision. For example if A were known to equal \$9,500, a supplementary purchase of \$5,500 would bring the stock up to the required total value of \$15,000. But if A were found to be \$17,450, the store already would be well stocked, and, in fact, overstocked. No further purchases would be necessary.

In general, A serves as the decision-parameter in the following way: *If A is less than \$15,000, the store will buy $(\$15,000 - A)$ dollars worth of jewelry; while if A is equal to or greater than \$15,000, they will make no further purchases.*

Note how this decision rule relates actions to observations before the observations are to be made. It implements the idea of a priori planning. Also make note of the fact that a statistical study, if it were to be used in such a problem, would aim at estimating A on the basis of sample information.

The last two illustrations serve to indicate only two of the many decision problems in which a population aggregate is the key decision-parameter. For problems in the field of accounting, information may be needed on *total* payroll, on the *aggregate* value of physical property, or on the sum *total* of checking account balances. Various economic problems require information on the economy's national income (the *aggregate* earnings of labor and property that arise from the current production of goods and services by the economy) or on the *total* amount of money in circulation or on *total* consumer credit. A purchasing agent may need to know the *total* backlog of unfilled orders in a supplier's industry before deciding on his own buying policy. A real estate corporation may need data on the *aggregate* amount of office space currently available in a city before making further construction plans. In all of these problems, and in many others as well, you can see that attention centers on some aggregate.

The aggregate is, of course, a reasonable property only for finite populations. The aggregate for any infinite set of variates ordinarily would be infinite. Thus A may be a decision-parameter only in enumerative studies, never in analytical studies.

In any problem in which A is the decision-parameter and it is known or it can be made known easily, the problem vanishes—no uncertainty about A would exist and a decision could be made in the face of complete information. If A is unknown, however, the problem may still be decided on the basis of a sample study; that is, partial information derived on the basis of statistical methods and theory could be used to estimate A or to otherwise provide a basis for decision.

In addition to π and A it will be convenient, at this time, to define and utilize some other statistical symbolism which we shall be using occasionally in the course of our studies.

First of all, the total number of observations in a population will be denoted by N . Thus N is a general symbol for the size of a population. Hence, in Example 3.8, since we were dealing with 20 variates, $N = 20$.

Secondly, each individual observation in a population will be denoted by X . However, in order to distinguish between observations we must add a subscript to each X . Thus, if a population is in some specified order, X_1 will refer to the first observation in the population, X_2 will refer to the second observation, X_3 to the third observation, etc. The last or N th observation will be denoted by X_N . For example, in Example 3.8:

$$X_1 = 1,225.04, X_2 = 218.96, X_3 = .00, \dots, X_N = X_{20} = 67.05.$$

(The three dots, incidentally, mean "etc.")

In terms of this notation, you can note that a population may be represented, in a general way, by the set of observations $(X_1, X_2, X_3, \dots, X_N)$. Now, as you know, a decision-parameter is some important property of a population. Alternately, we may say that a decision-parameter is some function of the N possible observations $(X_1, X_2, X_3, \dots, X_N)$.

The aggregate is a particularly simple decision-parameter; that is, it is a particularly simple function of the population $(X_1, X_2, X_3, \dots, X_N)$. It is merely the sum of these observations. Thus, since we have named the aggregate A , we may define it symbolically as:

$$A = X_1 + X_2 + X_3 + \dots + X_N.$$

One last refinement in this connection involves introducing an operational symbol for the operation of summing. Whenever we sum a series of numbers, we may symbolize that operation by Σ (the Greek capital letter *sigma*). Σ is similar to such operational symbols as $+$ which tells us to add, $\times$ which tells us to multiply, and $\div$ which tells us to divide.

To make use of the Σ , let us denote some value in a population by X_i —this refers to the i th observation in the population, where i is itself a general symbol which may take any value from 1 to N . If we want to express the operation of summing all of the X_i values, we may then write ΣX_i . This says "sum the values of X_i ." Which values are they? This depends on what we let i equal. For example,

$$\Sigma X_i \qquad (i = 1, 2, 3)$$

means $X_1 + X_2 + X_3$. Similarly,

$$\Sigma X_i \qquad (i = 2, 3, 4, 5, 6)$$

means $X_2 + X_3 + X_4 + X_5 + X_6$. Furthermore,

$$\Sigma X_i \qquad (i = 1, 2, \dots, N)$$

means sum everything; that is, it means $X_1 + X_2 + \dots + X_N$.

In conclusion, by using this summation symbol we may define the aggregate in a compact formula, namely,

$$A = \sum X_i \quad (i = 1, 2, \dots, N).$$

This expresses the thought that the population aggregate is the sum of all the values of X_i with i ranging from 1 to N .

3.4 The Arithmetic Mean

Another well-known and often-used idea—that of an “average”—is the basis for a third decision-parameter which frequently occurs in statistical decision problems. This decision-parameter is called the *arithmetic mean*.¹

Undoubtedly you are quite familiar with what most people call *the average*. For example, many students are quite adept at averaging their quiz grades—if one receives grades of 85, 100, and 100, usually he can calculate his average mark of 95 rather quickly, and even another with grades of 50, 50, and 60 can determine his average of $53\frac{1}{3}$ in due time. Thus, *the average* is, by custom, the sum of a set of values divided by the number of those values.

In statistics, however, *the average* most often is called the arithmetic mean. This name is preferred to the plainer term “average” for various reasons, including the fact that there are many measures called “averages” in statistics. We shall, for example, meet two other “averages” in the next section.

The arithmetic mean is used in many ways in and out of the field of statistics. For the present, however, we are interested in only one of these uses—the applicability of the arithmetic mean as a decision-parameter in statistical decision problems. In this section, then, we shall study the arithmetic mean as a population measure and consider the type of problem in which the mean adequately describes states of the world; that is, in which it represents the key fact about a population which a decision maker would like to know.

Example 3.10 A food products manufacturer purchases large quantities of an oil-bearing seed. These seeds are crushed for their oil, which is one of the products the company markets. The manufacturer periodically faces a decision problem on whether or not to accept a given shipment of seeds. In each case the average or mean number of pounds of oil per bushel may be adopted as an appropriate decision-parameter.

¹The term “average” is very old, dating back at least to the fifteenth century. The term “mean” is very probably even older and was known in the sixth century, B.C.

Suppose that for a particular shipment the cost is \$2.40 per bushel. Further assume that the value of a pound of oil is \$.30. If the average number of pounds of oil per bushel of seed were 10, the average value of a bushel would be \$.30 times 10, or \$3.00. Ignoring any other costs, it would be profitable for the firm to accept that shipment.

However, if the arithmetic mean of the amount of oil per bushel were only 7.5 pounds, the average value per bushel would be only \$2.25 (\$.30 $\times$ 7.5). The cost of \$2.40 would indicate clearly that acceptance of the shipment would be unprofitable.

The consequences of any decision depend on the value of the mean. It therefore represents a method of statistically describing states of the world for the problem.

However, the mean never could be made known exactly in this type of problem; at least not in time for the decision. Why? Nevertheless a well-selected sample may provide an alternative decision procedure.

Example 3.11 Among the many facts on which stock market investors base their financial decisions are the so-called "stock market averages." Not all of these averages are arithmetic means, but most of them are essentially of that form. Those that are may be interpreted quite simply.

A market average measures the average value of shares of stock in a well-defined portfolio of stocks. The definition of such a portfolio, although a difficult problem, is not, primarily, a statistical one. It is, in essence, an exercise in the definition of a population, and as such, the responsibility for determining the adequacy of any portfolio rests with an expert in finance.

For example, suppose a portfolio, deemed adequate for an investor's purposes, is defined as 25 shares of American Telephone and Telegraph Company, 100 shares of International Shoe Corporation, and 50 shares of Ford Motor Company. On a given day, if the closing prices of these stocks are \$175, \$33, and \$42, respectively, the total value of the portfolio would be

$$\$175(25) + \$33(100) + \$42(50) = \$9,775.$$

Hence, the average value per share is \$9,775 divided by the total number of shares (25 + 100 + 50 =) 175, namely

$$\frac{\$9,775}{175} = \$55.86$$

to the nearest cent.

Almost all stock market averages may be interpreted as the average value of some portfolio of stocks. As such any one presents an example of the arithmetic mean as a decision-parameter.

Example 3.12 The Revolutionary War brought with it a severe inflation, and the colonies' paper money apparently depreciated quite

rapidly. To offset the effect of the price rise, Massachusetts passed a law which provided for an *average of price relatives* to be used in determining the payment of wages to American soldiers from that colony.²

This scheme was designed to compensate for the rapid depreciation of the colonies' paper money during the war period by paying the Massachusetts soldiers in "depreciation notes." These notes were redeemable in amounts varying proportionately to the average of the relative prices of four commodities—corn, beef, wool, and sole leather.

Thus, "base period" prices were set for these commodities as their cost at the end of 1776; namely, in shillings/pence, as:

Beef (lb.)	/3½	Wool (lb.)	2/0
Corn (bu.)	4/0	Leather (lb.)	1/3

Furthermore, "current" prices were collected monthly by a group of about one dozen prominent citizens, who had been appointed for that task. Thus, in June, 1777, the following prices (again in shillings/pence) were reported:

Beef (lb.)	/8	Wool (lb.)	4/0
Corn (bu.)	6/0	Leather (lb.)	1/3

Based upon these data, price relatives were computed for each of these commodities. That is, the ratio of each current price to the corresponding base period price was computed. Thus, for June, 1777, the price relative for corn was 6 divided by 4 (shillings), or 1.500; in fact, as reported by the prominent citizens, the 4 price relatives were

Beef (lb.)	2.284	Wool (lb.)	2.000
Corn (bu.)	1.500	Leather (lb.)	1.000

and their average, the depreciation rate,

$$\frac{2.284 + 1.500 + 2.000 + 1.000}{4} = 1.69 .$$

(You might check their arithmetic!)

This result indicates that prices rose 69 per cent between the end of 1776 (the base period) and June, 1777, at least as measured by this average. In any event, the average served as a basis for the decision on how much to pay Massachusetts' soldiers. For example, in June, 1777, a depreciation note issued at the end of 1776 was redeemable at 169 per cent of its face value.

Today economists call averages of price relatives *index numbers*. Roughly speaking, index numbers measure over-all changes in the level of prices (or in the level of some other economic concept). They often are used as decision-parameters today in much the same way as in our example from 180 years ago. Thus, the wages of over 4 million blue-collar workers in the United States are tied, by contract, to

² Willard C. Fisher, "The Tabular Standard in Massachusetts History," *The Quarterly Journal of Economics*, Vol. XXVII (May, 1913), pp. 417-54.

the Consumer Price Index, an index number of prices (for about 300 items) published by the U.S. Bureau of Labor Statistics.

As noted earlier, some statistical problems involve not one but two or more decision-parameters. Thus, at times, the average alone is not sufficient information about a population unless supplemented by other criteria.

Example 3.13 A foundry produces iron castings. It must decide, hourly, whether or not its manufacturing process is operating satisfactorily. If so, it takes no action; if not, it looks for trouble in the process, and if found, the trouble is corrected. An elementary unit is a casting; a characteristic of interest, its weight. The population is the totality of weights the foundry's process could turn out if it operated indefinitely under substantially the same operating conditions. What are the decision-parameters?

One is the arithmetic mean of the weights—this obviously is important for if the castings are too light, on the average, the foundry is not giving customers enough for their money. However, the arithmetic mean, in itself, does not describe states of the world completely. Another important consideration in the problem is that if a casting is too heavy (e.g., over 22.5 pounds), it cannot be processed further by some customers. Hence, the proportion of castings over 22.5 pounds is a second important decision-parameter.

In short, the decision problem involves at least two decision-parameters: the mean and a proportion. The arithmetic mean, though important, is not sufficient information in itself to reach a decision.

The arithmetic mean of a population may be represented, symbolically, as μ (the Greek letter *mu*). Thus, μ is defined, for a population of variates, as the sum of all the variates divided by the number of them.

In many problems the arithmetic mean (μ) may be used interchangeably with the aggregate (A) as a basis for decision. As you may recall, N denotes the number of observations in a population, and hence the arithmetic mean is simply the aggregate divided by N ; that is, the arithmetic mean and the aggregate are related in the following ways:

$$\mu = \frac{A}{N} \quad \text{and} \quad A = N\mu.$$

In several problems, however, it is advantageous to deal in terms of the mean rather than the aggregate. For one thing, the procedure of taking an average reduces the aggregate to a per unit figure. Thus, even if we knew that the total annual payroll in one company is \$1.5 million and in another it is \$1.2 million, the figures might be

misleading until we knew the average wage per worker or per some other elementary unit. This, however, depends on the problem situation.

For another, the aggregate has no meaning in analytical studies since, as noted, the sum of an infinite number of variates ordinarily is infinite. But an infinite population still may have an arithmetic mean. For example, to determine coffee bean requirements we might like to know the average (arithmetic mean) weight of coffee we are packaging in "pound" cans. Such a study is analytical, dealing as it does with a process—hence the population of net weights is infinite. It represents all possible weights of coffee the process would turn out if it continued to operate indefinitely under the same conditions.

We may define the arithmetic mean symbolically in terms of the notation introduced in the last section; that is, as a function of the N possible observations in a population: $(X_1, X_2, X_3, \dots, X_N)$.

Thus,

$$\mu = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N}$$

or, more compactly, by reintroducing the summation symbol

$$\mu = \frac{\Sigma X_i}{N} \quad (i = 1, 2, \dots, N).$$

Since ΣX_i with $i = 1, 2, \dots, N$ is the sum of all the variates in a population, μ is defined as it was earlier—the sum of all the variates divided by the number of them.

The arithmetic mean has several interesting arithmetical properties, two of which we may note briefly here. Proofs and/or further examples are left to your curiosity and discretion.

1. Given a set of N numbers with a mean μ : if another number less than μ is added to the set, the new mean is less than μ . For example, the mean of 2, 3, and 4 is 3; if we add 1 to our original set, the mean decreases to 2.5. Similarly, if a number greater than μ is added to a set of numbers, the mean increases. What happens if a number equal to the original mean is added to a set of numbers?
2. The sum of the differences between each individual value and the arithmetic mean invariably equals 0. For example, the mean of the numbers 3, 4, and 8 is 5. The differences or "deviations" from the mean are $3 - 5 = -2$, $4 - 5 = -1$, and $8 - 5 = +3$, and the sum of -2 , -1 , and $+3$ is 0. Symbolically this property may be expressed as

$$\Sigma(X_i - \mu) = 0 \quad (i = 1, 2, \dots, N).$$

3.5 The Median and the Mode

It is a most amazing fact that so many business decision problems can be solved by information about a population proportion, or an aggregate, or an arithmetic mean. These three decision-parameters are simple yet useful in a large variety of situations, and, hence, a good part of our introduction to statistics will center on them.

At this point, however, we may make note of the fact that there exist many other decision-parameters which occasionally represent pertinent facts about a population of study and briefly reinforce this assertion with a discussion of two such measures. These two measures are called the *median* and the *mode*.³

Let's consider the median first. The median of any given set of numbers may be described as the middle or central value in that set of numbers. Or it may be described as that value above which and below which half of the set of numbers lie. Thus, we may say, for the moment, that when any set of numbers is arranged in order of magnitude, the middle value is referred to as the median. Suppose we look at a numerical illustration or two before trying to define the median more precisely.

Suppose the straight-time hourly wages of five workers are \$2.10, \$2.08, \$2.15, \$2.00, and \$2.12. To determine the median we must arrange these five values in their order of magnitude—\$2.00, \$2.08, \$2.10, \$2.12, \$2.15. Then the median is \$2.10 since that is the middle value. Such a "middle value" always exists when we are dealing with an odd number of values.

However, suppose we want to derive the median for an even number of values. For example, the monthly rents for six apartments are \$128, \$193, \$150, \$187, \$162, and \$96. To find the median rent we array these six values—\$96, \$128, \$150, \$162, \$187, \$193. What is the median?

It should be the middle value, but unfortunately there is no middle value; there are two, namely, \$150 and \$162. In any such case, either of the two middle values or any number between them may be

³ The names "median" and "mode" are relatively recent. Sir Francis Galton (1822-1911), outstanding early statistician, biologist, mathematician, and cousin of Charles Darwin, developed the idea of a median in the latter part of the nineteenth century.

Karl Pearson (1857-1936) who introduced the term "mode" may fairly be called the founder of statistics. Any one- or two-sentence comment on Pearson here would be inadequate, but the interested student might read articles about him by Helen M. Walker and Samuel A. Stouffer in the March, 1958, issue of the *Journal of the American Statistical Association*.

denoted as the median. However, for convenience, statisticians usually adopt the convention of defining the median as the mean of the two central values. Hence, under this convention, the mean of \$150 and \$162, namely, \$156, is the median of the above rents.

A precise definition of the median of a set of numbers is as follows: The median is a value that is neither greater than more than half of the numbers nor less than more than half of the numbers.

Thus, the median of an odd number of values is, by definition, always the middle value when the complete set of numbers is arrayed. The median of an even number of values is, by convention, the mean of the two middle values.

In what types of problems is the median useful as a decision-parameter? The answer to this question is standard—any problem in which it is a pertinent fact which a decision maker would like to know. However, to better recognize its ability to serve as a decision-parameter, it is helpful to know certain properties of the median.

As was implied earlier, the median generally is regarded as an average—not *the* average, which is the term laymen reserve for the arithmetic mean, but simply as one of several averages. Like the mean, the median summarizes a population in one single, “typical” number. It is natural, then, to compare the median to the arithmetic mean.

Every population of variates has a mean and a median as well. In some populations the values of these two measures are equal; in others the mean is larger than the median; and in still others the median is larger than the mean. One reason for this rests with the fact that the mean is influenced by extreme values while the median is not. For example, suppose we have three families’ annual incomes—\$3,000, \$4,000, and \$38,000. The mean of these incomes is \$15,000; the median is \$4,000. The mean is larger than the median because the mean is more affected by the one large, extreme income of \$38,000.

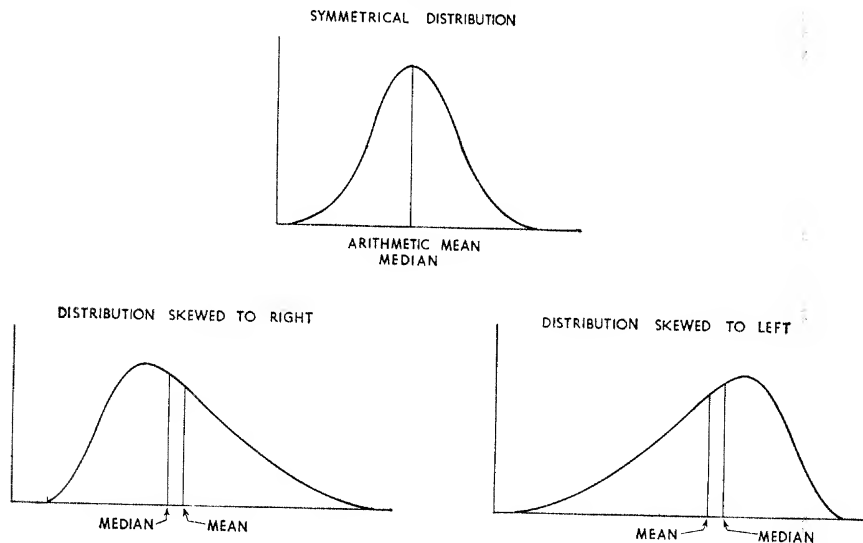
Example 3.14 One point of conflict in labor-management wage negotiations is the appropriate decision-parameter with which to measure average wages and upon which to negotiate. One of management’s chief interests is total costs, hence it focuses on the arithmetic mean—the mean wage multiplied by the number of employees equals total payroll. A union, however, usually is more interested in the median wage since it is a better criterion to determine the welfare effect of a wage change on its members; 50 per cent make the median wage or more and 50 per cent make the median wage or less.

This usually presents a conflict situation, for the mean (upon

which management focuses) is often considerably larger than the median (upon which the union focuses) because the mean is more affected by a few large, extreme wage rates for highly skilled workers.

In general an answer to the question of whether the mean or the median of a population is the larger, or whether the two are equal, depends on how the variates in the population are distributed. To see this graphically, consider Figure 3.2 which shows three population models. The first is symmetrical, and in any such population the

Figure 3.2



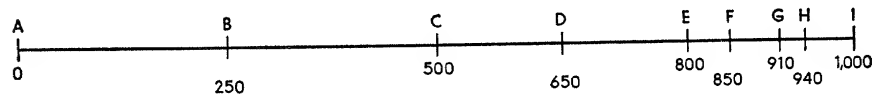
mean and the median are equal in value. The second is skewed to the right, and in such a case, as shown, the mean is greater than the median. Why? Because the very fact that the curve tails off to the right means that the population includes some high extreme values. These tend, as they should, to inflate the mean, yet have little influence on the median. The third distribution shown is skewed to the left. This implies some low extreme values, and hence the mean is smaller than the median. Figure 3.2 illustrates the fact that knowledge of the general shape of a population—that is, the population model—may be of some aid in determining the relationship of the mean and the median.

Thus, the fact that the mean is affected by extreme values in a population while the median is not, points up a general principle to follow in choosing one measure or the other as a decision-parameter: The median is preferable to the mean if you do not want your deci-

sion to be affected by extreme values. Of course, for some problems the very fact that the arithmetic mean is affected by extreme values in relation to their magnitude and to their frequency of occurrence makes it the proper decision-parameter.

Another property of the median which is of some interest arises in connection with the absolute differences between the median and the variates in the population it represents. (An absolute difference is one in which the sign, plus or minus, is ignored.) The property of interest is that the sum of the absolute differences between the median and the individual variates in a population is a minimum; that is, smaller than any sum of the differences between the variates and any other value. For example, consider an array of numbers: 9, 10, and 15. The median is 10. The absolute differences between the original numbers and the median are 1, 0, and 5, respectively, and this totals 6. This is a minimum; that is, the absolute differences of 9, 10, and 15 from any value other than the median will add up to more than 6. (You might check this by measuring absolute deviations about a value other than the median.)

Example 3.15 A concessionaire has nine refreshment stands located at various points on the boardwalk of a popular beach. He wants to locate a central supply station at some point on the boardwalk, and he wishes to locate the supply depot so that the amount of travel between the stands and the source of supply is minimized. Graphically, we may envision the stands and their distance in yards from stand A as follows:



We have introduced a population of observations (distances) into the problem. Furthermore, the decision-parameter of interest is the median, for this point will minimize the total distance to all other points of interest. Thus, the median is the fifth observation, or 800, which represents stand E.

The concessionaire might locate at this point. In doing so he will minimize the total distance from the supply station to the stands. This total distance is, incidentally,

$$(800 - 0) + (800 - 250) + (800 - 500) + \dots + (1,000 - 800) = 2,280 \text{ yards.}$$

Any other location which might be chosen as the supply depot would yield a larger total distance.

However, it should be noted that the simple solution presented rests on the assumption that each stand would be supplied equally

often. If not, we would have to count each observation (distance) in our population in proportion to the frequency with which it requires supplies, and determine the median of that population.

For example, you can see intuitively that if A required supplies far more often than any of the other points, total distance would be minimized not at E (which no longer would be the median) but at some point closer to A. Whatever that point, however, it would be the median of the population of distances included in proportion to their occurrence.

Now consider a decision-parameter called the mode. The mode of a population may be defined as the most frequently occurring variate (or attribute) in that population. Consider, as a simple illustration, a set of five numbers: 2, 3, 3, 3, and 6. The mode is 3 because it occurs more frequently than any other number.

Graphically, the mode has a simple interpretation. Consider Figure 3.2. For each population model, the peak or the highest point, which implies the most frequently occurring value, indicates the position of the mode.

The mode, like any other property of a population, is an appropriate decision-parameter in situations when it is a key bit of information which a decision maker would like to know. While such situations do not present themselves too often, there are exceptions. Generally, these involve a decision problem in which the most common variate (or attribute) in a population must be selected as a basis for action.

Example 3.16 Stores that sell at a low markup depend on a large turnover to make a profit on their operations. They can accomplish this by carrying the most typical or modal size of merchandise. A fast turnover would be defeated by a policy of carrying many sizes. Hence, they carry that size which meets the demand of the largest number of people. This is, by definition, the mode. Therefore, in order to decide on what merchandise to stock, the mode is the decision-parameter of interest.

Note the difference in objectives between the stores who buy for the modal customer and others, such as the haberdasher in Example 3.3, who buy for the whole distribution of customers.

3.6 Variability in Populations

As has been pointed out on one or two occasions, populations are characterized, almost invariably, by some *variability*—by some differences in the observations that make up a population. Your statistics quiz grades vary from test to test. Wage rates differ from occupation to occupation. Prices differ from product to product. The

number of persons reading *The New York Times* varies from day to day. Even the diameters of ball bearings differ from one bearing to another—although a producer hopes that this variation is not great.

In this section, and the next as well, we shall study this idea of variability in populations, how it is useful in statistical decision problems, and the ways in which we can be more specific about the notion of variability itself. To start being more specific, suppose we consider a small population of five wage rates: \$2.20, \$2.60, \$2.10, \$2.25, and \$2.35. This simple population obviously is characterized by some variation—the wage rates differ. But how much variation is there in this set of numbers? This is one of the basic questions with which we must now deal.

Unfortunately, it is a difficult question. How, to consider a similar question, do you measure the consistency, or lack of consistency, in your quiz grades? One student gets grades of 50, 55, 50, and 60. We say, among other things, that he is relatively consistent. Another receives quiz marks of 70, 100, 65, and 95. He is relatively inconsistent. Still another gets grades of 95, 95, 95, and 100. He is almost perfectly consistent. But what do we mean by consistency? Is there any way of measuring the consistency of, or the variation in, a set of numbers (e.g., wage rates or quiz grades)?

The answer to this question is: "Yes, there is a way of measuring variability." In fact there are several ways, three of which we will consider in this section. An analogy with some of our previous work also may help in pointing up where we are headed. The arithmetic mean, the median, and the mode, which we have studied in the last two sections, present numerical measures of an "average" for a population. Now we are interested in a numerical measure of variation in a population. In the first case we are concerned with a single number which represents an average value for a population. Analogously, in the second case we want to define a procedure for arriving at a single number which represents the variation in the population.

Before we attempt to define any such measure of variation, it is reasonable to ask why we should bother doing so. There are several reasons, one of which we will consider now. This involves the fact that the consequences of decisions in some problems depend on the variation in a population of study. In such cases we need an appropriate measure of the variation to describe states of the world. Such a measure, if known, would be the pertinent fact a decision maker needs in order to choose among his alternatives. Thus, in short, one reason we want a measure of variation for populations is so that it

will serve as a decision-parameter. Let's look at some illustrations in which variability is an important criterion for making a decision.

Example 3.17 A common decision problem in production control occurs in choosing between two (or more) ways of manufacturing the same item; that is, in selecting that process which is capable of producing the better (or best) results. What does better (or best) mean? It may have several interpretations. One important one, which is pertinent to our present discussion, flows from the fact that one of the main objectives of every manufacturing company is to produce a relatively uniform product.

Thus, with respect to this matter of uniformity, or alternately that of variation, one process is better than another if it produces a product which is more uniform, or less variable. Consequently, one decision-parameter of interest is some measure of the variation in the populations associated with the processes on which the decision is to be made.

Example 3.18 Jones and Smith both invest in the stock market. They have, however, different objectives. Jones' goals are typical of a large class of people who aim for a steady dividend income and who want to take little risk. Jones and similar investors are interested in stocks whose dividends promise little variation. Smith's goals are different. He is a speculator and is interested in those stocks whose prices promise considerable variation. Thus, variation, or the lack of it, is one important criterion for Smith or Jones in deciding whether or not to purchase a given security.

Example 3.19 Until recent years, ladies' coats of a given size were all manufactured and sold with uniform lengths. Firms that manufactured coats used only some average length as their decision-parameter. Short girls, upon buying a coat, had to cut off material or otherwise shorten it. Tall girls had to let down a hem to lengthen their garment. In addition, manufacturers had to use enough material on each garment to meet the length requirements of tall girls.

Recently, however, the industry has begun to take *variability* into consideration in the production of ladies' coats. Every coat size is now made in three lengths: short, regular, and long. For the manufacturer, using variability as a criterion for decision means reduced costs on short and regular garments; while for many consumers it means a better fitting garment.

Thus, we have sufficient motivation for defining a numerical measure of variation. But the problem of exactly how to define such a measure is still a difficult one. Suppose we begin to overcome this difficulty with a simple, though not altogether satisfactory, solution.

The Range. Such a solution is provided by a measure called the *range*. The range of a population is defined as the maximum value less the minimum value. Thus, for a population of wage rates, such as \$2.20, \$2.60, \$2.10, \$2.25, \$2.35, the range is \$.50 (the maximum value, \$2.60, less the minimum \$2.10).

Unfortunately, despite its simplicity, the range is not too useful as a decision-parameter. Its value depends only on the two extreme variates in a population; it in no way expresses the variation of the other variates lying between these two extremes. Ordinarily the consequences in a decision problem depend on some over-all measure of variability rather than on the range alone. Thus, the range is not particularly useful as a decision-parameter. It does have, however, other uses in statistics.

The Average Deviation. Since the range has limitations as a measure of variability in a population, suppose we consider another possible approach. Note first that if all observations in a population were identical, there would be no variability. Furthermore, in such a case there would be no difference between any one value and the arithmetic mean of the observations since the mean would be identical with each and every one of the individual observations. However, the more variable a population, the greater would be the differences between the mean and the individual observations.

Suppose, then, we focus on these differences or deviations of the individual observations from their arithmetic mean. Specifically, let's once again consider our small population of wage rates: \$2.20, \$2.60, \$2.10, \$2.25, and \$2.35. The arithmetic mean of this population is easily determined as follows:

$$\mu = \frac{\$2.20 + \$2.60 + \$2.10 + \$2.25 + \$2.35}{5} = \frac{\$11.50}{5} = \$2.30.$$

Therefore, the deviations on which we are interested in centering our attention are as follows:

$$\begin{aligned} \$2.20 - \$2.30 &= -\$0.10 \\ \$2.60 - \$2.30 &= +\$0.30 \\ \$2.10 - \$2.30 &= -\$0.20 \\ \$2.25 - \$2.30 &= -\$0.05 \\ \$2.35 - \$2.30 &= +\$0.05 \end{aligned}$$

We will consider two measures of variability which are based on deviations such as these.

First, however, note that it would be futile to attempt to define a measure of variation as the average of these deviations because

the sum of the deviations about the arithmetic mean is always equal to zero. This is true in our simple numerical example since

$$- \$.10 + \$.30 - \$.20 - \$.05 + \$.05 = 0 ,$$

and it is true for any and every set of numbers, as was pointed out in Section 3.4.

The fact that some deviations are positive and some are negative causes our trouble, for the positive total is canceled out by the total of the negatives. Suppose, however, that we ignore the signs of the deviations and then take their arithmetic mean. That is, suppose we define a measure of variation as the average of the *absolute* deviations of observations from their arithmetic mean. For example, the *average deviation* (or *mean deviation*) for our wage rate data is:

$$\frac{\$.10 + \$.30 + \$.20 + \$.05 + \$.05}{5} = \frac{\$.70}{5} = \$.14 .$$

The average deviation is, obviously, a reasonable measure of variability and rather simple to interpret as well. It could, in many problems, be adopted as a decision-parameter. However, this measure is difficult to deal with mathematically because of the "unnatural" algebraic procedure of ignoring the signs of the deviations. Furthermore, it is of little use in sampling theory.

The Standard Deviation. Because of this fact and also because of historical reasons, statisticians often employ another measure of variation, also based on deviations about the arithmetic mean.

The difficulty in basing a measure of variation on the deviations about the mean lies with the fact that the negative deviations cancel out the positive deviations. The average deviation makes use of one remedy—ignoring the signs of the deviations. Another solution to the dilemma is to base a measure of variation on the squares of the deviations. Once again the negative signs on the deviations will be eliminated.

Thus, in the simple numerical situation we have been discussing throughout this section, the population of five wage rates is:

$$\$2.20, \$2.60, \$2.10, \$2.25, \$2.35 .$$

In view of the fact the mean of these rates is \$2.30, the deviations about the mean are:

$$- .10, + .30, - .20, - .05, + .05 ,$$

and if we square these deviations, we find

$$.0100, .0900, .0400, .0025, .0025 .$$

The average of squared deviations such as these is called the *variance*. Thus, for our numerical example, the variance equals

$$\frac{.0100 + .0900 + .0400 + .0025 + .0025}{5} = \frac{.1450}{5} = .0290 .$$

One other step usually is convenient in dealing with a measure of variation. Because we averaged *squared* deviations, our answer is, so to speak, in the wrong unit of measurement (e.g., dollars squared!). Therefore, it is often easier to deal in terms of the square root of the variance. This is called the *standard deviation*. For our numerical example, the variance is .0290; hence, the standard deviation is:⁴

$$\sqrt{.0290} = \$.17 .$$

Throughout our studies of statistics we will have several uses for the variance and the standard deviation.⁵ These uses include their applicability as decision-parameters and, more generally, as measures of variability in a population. We will prefer these measures to the range and the average deviation for reasons which include those already noted. Remember, then, that the variance is the average of squared deviations from the arithmetic mean, while the standard deviation simply represents the square root of the variance. Sometimes it will be easier to deal in terms of the variance; other times to deal in terms of the standard deviation.

Because we will be referring to them frequently, it is convenient to give the variance and the standard deviation symbolic names. Thus, the standard deviation of a population will be called σ (the Greek lower-case letter *sigma*) and the variance σ^2 (*sigma* squared). Thus, for our wage rate illustration

$$\sigma^2 = .0290 \text{ and } \sigma = \sqrt{.0290} = \$.17 .$$

The definition of the variance may be put precisely in terms of the symbolism we have introduced throughout this chapter. Given a population of N observations, $(X_1, X_2, \dots, X_N)$, we consider the differences between each of these values and their arithmetic mean, μ :

$$(X_1 - \mu, X_2 - \mu, \dots, X_N - \mu) .$$

⁴ You will find Appendix A of some use whenever you need to compute the standard deviation and/or the variance. It includes a table of squares and square roots, and instruction on how to use it.

⁵ The term "standard deviation" was introduced by Karl Pearson in 1893. The term "variance" was first used in 1918 by another renowned statistician, Sir Ronald Fisher.

The variance is based on the squares of each of these deviations from the mean; that is, on $(X_1 - \mu)^2$, $(X_2 - \mu)^2$, . . . , $(X_N - \mu)^2$. Specifically, the variance is defined as

$$\sigma^2 = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_N - \mu)^2}{N},$$

or, making use of the summation sign, Σ , we may write

$$\sigma^2 = \frac{\Sigma(X_i - \mu)^2}{N} \quad (i = 1, 2, \dots, N).$$

The standard deviation may be expressed in a similar way. Since it is simply the square root of the variance,

$$\sigma = \sqrt{\frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + \dots + (X_N - \mu)^2}{N}}$$

or, equivalently

$$\sigma = \sqrt{\frac{\Sigma(X_i - \mu)^2}{N}} \quad (i = 1, 2, \dots, N).$$

How do you interpret the variance and the standard deviation? Since this question arises so often, perhaps we had better devote the next section completely to answering it.

3.7 The Standard Deviation and the Normal Curve

You should recognize that the variance (as well as its square root, the standard deviation) holds a position of considerable importance in the development of statistical methods. Although to some extent the choice of that particular measure of variability seemingly is arbitrary, the choice is well founded in theory. In any event, the selection of a particular measure is of secondary importance as compared to its proper utilization and interpretation. Suppose, then, in this section we attempt more precisely to interpret σ^2 (and σ).

The first thing for you to remember, generally speaking, is that σ^2 , or σ , measures in one over-all number the variability in a population. It is a summary indicator of that particular population characteristic. The greater the variability in the population, the larger is the value of σ^2 or σ . Secondly, you must note that most measures of variability, including the variance, represent a measure of the extent to which there are differences between individual observations and some central or average value. This is precisely what variation means—departure from a norm. In the case of σ^2 , as we have noted, differences from the arithmetic mean of a population are involved in its definition.

To interpret a given value of σ more precisely, one ordinarily must think in terms of some population model. Thus, in many problems, if we assume a certain population model, we may deduce the relative frequencies of observations that fall within one, or two, or any given number of standard deviations from the arithmetic mean of that population. In order to illustrate this point in a concrete way, we may reconsider a population model briefly touched on earlier; the *normal curve*.

The normal curve apparently was first discovered by a French mathematician, Abraham De Moivre,⁶ in 1733, in connection with his studies of games of chance. Over the next 50 years his discovery went unnoticed. Meanwhile several other mathematicians independently derived the normal curve. One of these rediscoverers was the German mathematician, Karl Gauss.⁷ In the subsequent development of mathematical statistics, De Moivre's work was forgotten, and the theory of the normal distribution stemmed from the work done by Gauss. For this reason the normal curve is sometimes referred to as the Gaussian distribution.

The work of Gauss and his contemporaries led to a wide applicability being attributed to the normal curve. Thus, in the early days of its development many people thought that most statistical populations could be approximated by this population model. For that reason they attached great importance to the normal curve. Today, however, with considerably more experience behind them, statisticians are reluctant to make the generalization that most populations can be assumed normal. But, even so, many populations can be assumed normal, and, furthermore, the normal curve has several other uses in the field of statistics. Thus, the normal curve holds a central place in statistical theory.

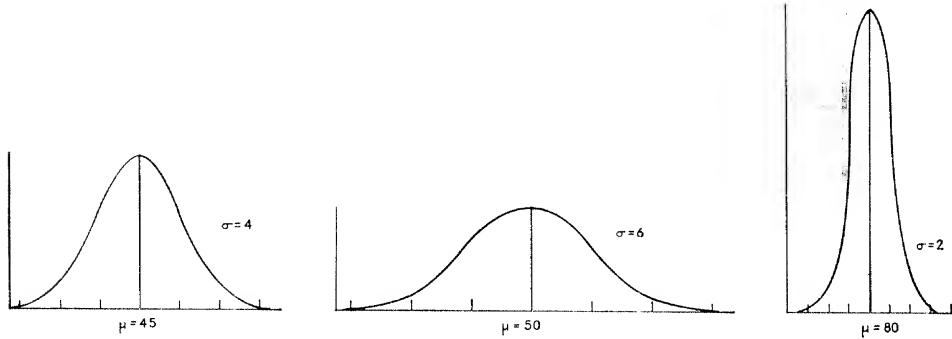
The normal curve may be expressed in terms of a mathematical formula. For our purposes, however, we can gain an understanding of this curve without resorting to its explicit formula. An analogy with the straight line is rather helpful. As you undoubtedly know, a straight line is determined by two points. A different pair of points

⁶ Abraham De Moivre (1667–1754) lived in England most of his life tutoring in mathematics and solving mathematical puzzles and problems for wealthy patrons. Among the questions for whose answers such patrons were willing to pay were those dealing with games of chance, i.e., cards, roulette, dice, etc. In this connection, De Moivre became one of the originators and developers of mathematical probability.

⁷ Karl Frederick Gauss (1777–1855), a famous German mathematician, contributed to almost every known branch of mathematics. His interest in the normal curve and in fact his derivation of the curve and its various properties were stimulated by his work in the field of astronomy.

generally yields a different straight line. Similarly, a normal curve is determined by two characteristics—its arithmetic mean and its standard deviation. Thus, if we pick an arithmetic mean and a standard deviation, we determine one normal curve; pick another μ and σ and we determine another normal curve. Generally, each different pair of these characteristics yields a unique normal curve. For example, Figure 3.3 shows us three different normal curves.

Figure 3.3



Notice that the centers of these normal curves all lie at different locations on their respective scales—their positions are governed by the values of μ . In addition, some of the curves are fatter (more variable) or skinnier (less variable) than others—this is governed by the value of the standard deviation because σ measures variability. As we said, then, a normal curve is determined by these two characteristics.

Ordinarily, a normal curve looks like a bell. At least, it will always look like a bell if you choose the scales on your graph appropriately or, alternatively, if you use your imagination vividly. In any event, it is from this fact that the alternative name “bell-shaped” curve obtains.

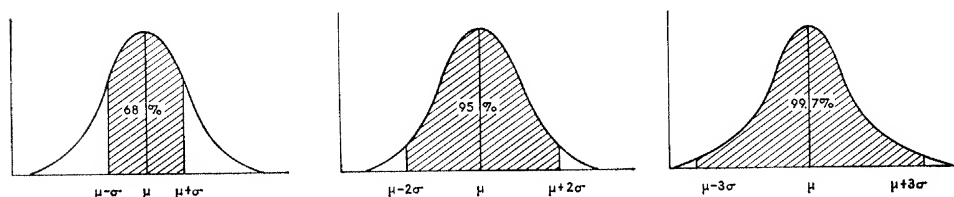
Nevertheless, every normal curve, whether it appears short and wide or tall and thin in a chart, has certain basic properties. For one thing, a normal curve is always symmetrical about its arithmetic mean. This means that the right half of the curve is a mirror image of the left half, and vice versa. Another interesting fact is that the arithmetic mean lies at the point at which the curve hits its maximum height. These remarks reflect the fact that the arithmetic mean, the median, and the mode of a normal population model are all equal. Why?

It also should be pointed out that the two "tails" of the normal curve trail off indefinitely in both directions. Thus, the range of a normal population is infinite. That is, both sides of the curve get closer and closer to the horizontal axis as we go farther and farther away from the center, but they never quite touch. This is a significant idea, and you should try to visualize it. Unfortunately, graphs cannot do justice to this point because the tails come so close to the horizontal axis that they appear to touch it.

In addition, it is very pertinent to any interpretation of σ for the normal curve to note that about 68 per cent of the area of *every* normal curve is included within the interval: (1) the arithmetic mean less one standard deviation, and (2) the arithmetic mean plus one standard deviation.

Thus, suppose we have a population that can be represented by the normal model. Approximately 68 per cent of the observations would

Figure 3.4



have values between $(\mu - \sigma)$ and $(\mu + \sigma)$. Of course, the actual magnitude of this interval depends on the values of μ and σ and, hence, on the particular population under consideration. Thus, of those normal curves in Figure 3.3, 68 per cent of the observations in population (a) are between 41 and 49 since $\mu = 45$ and $\sigma = 4$; 68 per cent of the observations in population (b) are between 44 and 56 since $\mu = 50$ and $\sigma = 6$; and 68 per cent of the observations in population (c) are between 78 and 82 since $\mu = 80$ and $\sigma = 2$.

Other relative frequencies besides the 68 per cent may be deduced for the normal curve. Thus, it can be shown that approximately 95 per cent of the observations lie within the interval represented by the mean plus and minus two standard deviations. Almost all (about 99.7 per cent) fall within the interval represented by the mean plus and minus three standard deviations. These relative frequencies hold for *every* normal curve. The ideas are shown graphically in Figure 3.4.

Example 3.20 A company manufactures floor tiles. Experience has shown that the thicknesses of the tiles that its production process turns out may be assumed to be normally distributed. Suppose it were determined that the mean of this population is .25 inches and that the standard deviation is .005 inches. It would follow that approximately 68 per cent of the floor tiles the process would produce, if it continued to operate under substantially the same conditions, would measure between .245 ($.25 - .005$) and .255 ($.25 + .005$) inches. Similarly, approximately 95 per cent of the tiles would have a thickness of between .24 and .26 inches, and about 99.7 per cent would be between .235 and .265 inches.

It will help you if you will recall that in any graphic representation of a frequency distribution, relative frequency is proportional to the area under the curve. Thus, the results noted above and those shown in Figure 3.4 are derived by measuring areas under the normal curve. These areas or relative frequencies are determined by advanced mathematical techniques. Fortunately, simple tables of the results are available, and such a table is given, together with directions for use, in Appendix B. We frequently will refer to the idea of relative frequencies or areas under the normal curve and will make extensive use of Appendix B.

For the present, however, note that the interpretation of σ may be facilitated by reference to a population model. If our population may be represented (approximately) by the normal curve, we know (approximately) 68 per cent of the observations in the population have values between $(\mu - \sigma)$ and $(\mu + \sigma)$, etc.

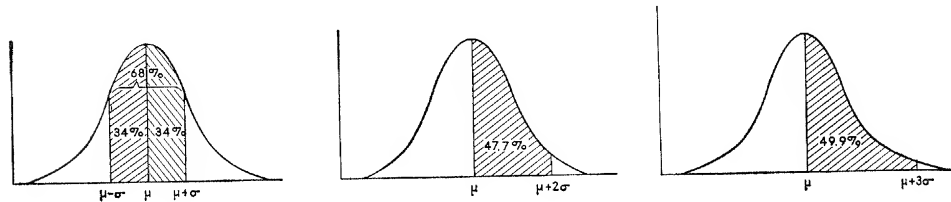
Of course, not every population may be represented by the normal model. Thus, this interpretation of σ , in terms of the relative frequency of the observations included in a given interval, depends on the particular model that is being utilized. Figure 3.5 points this up by a comparison of the normal curve with two other smoothed frequency distributions. They are called the exponential population model and the rectangular population model. The exponential model, being skewed, presents different problems of interpretation. For example, approximately 63 per cent of the observations in such a population fall in the interval one standard deviation to the left of the mean, while only about 23 per cent fall in the same interval to the right of the mean. But note in particular how the relative frequencies between, say, the mean plus or minus one standard deviation differ from model to model. In the case of the normal curve 68 per cent of the observations are included in that interval; in the case of the

exponential the relative frequency is 86 per cent; and for the rectangular model it is only 58 per cent.

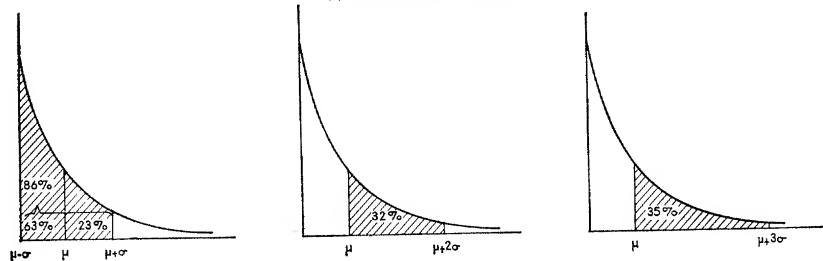
One last point in this connection. Suppose no population model is being used because none is applicable to the problem. Can we then make any statement about the relative frequencies that will fall

Figure 3.5

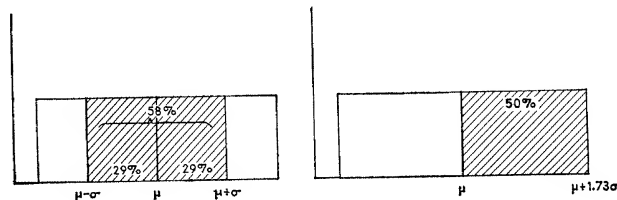
a. NORMAL DISTRIBUTION



b. EXPONENTIAL DISTRIBUTION



c. RECTANGULAR DISTRIBUTION



within the limits set by the mean plus and minus a given number of standard deviations? No exact answer can be given to this question. However, on the basis of statistical theory, we can say that the relative frequency of the observations in a population that will deviate from the mean by more than k standard deviations is equal to or less than $1/k^2$. For example, take $k = 2$. It follows that no more than $1/4$ of the observations in a population will be outside of the interval $(\mu - 2\sigma)$ to $(\mu + 2\sigma)$. At least 75 per cent of the observations will be within that range. Similarly with $k = 3$ no more than $1/9$ of the ob-

servations will fall outside of $(\mu \pm 3\sigma)$; at least 88.9 per cent must be within that range.

The general statement given above, known as Tchebycheff's⁸ inequality, is an example of what statisticians call nonparametric statistics. Roughly speaking, any theoretical result which does not hinge on the assumption of a population model falls within this branch of our subject.

3.8 Selecting the Decision-Parameter

In the preceding sections we have defined, studied, and illustrated several decision-parameters which occur rather frequently in statistical decision problems. The choice of some criterion such as one of these is a basic step in the planning stage of a statistical study—it focuses attention on what a decision maker wants to know about the population he has defined.

Judgment plays an important role in specifying the decision-parameter. As noted earlier the basic question that must be answered, of course, is: "What do we have to know about the population in order to make a decision?" Often, this is a simple question that can be answered readily. At times, however, it is a surprisingly difficult one and requires considerable deliberation. In any event, the choice of a decision-parameter calls for the combined judgments of management and the statistician.

This, however, is not to say that these judgments should be made arbitrarily or too subjectively. The choice of a decision-parameter should follow certain patterns of logical thinking. Thus, whatever the choice, the decision-parameter should adequately describe states of the world. It should be such that the consequences of a decision depend, at least in part, on the value of that decision-parameter. Only in these ways will knowledge of that decision-parameter provide a basis for rational action. Judgment plays an important role only in determining what is adequate, what may be the consequences, and how knowledge of the value of the decision-parameter would be useful in deciding on a course of action. In short, the choice of the decision-parameter depends on the particular circumstances surrounding the problem.

Example 3.21 An electronics company receives shipments of a special type of dry-cell battery. Shipments include about 500 bat-

⁸One of the great unsolved problems of mathematics and statistics is how to spell Tchebycheff's name. Some authors use Tchebycheff, others Chebycheff. In any event, Pafnutii Lvovitch Tchebycheff (1821-94) was a celebrated Russian professor of mathematics at the University of St. Petersburg during the nineteenth century.

teries, and for each shipment the decision problem is whether to accept or to return the batteries to the supplier. What is the appropriate decision-parameter by which to measure quality of the lot—that is, by which to describe states of the world? This depends on the circumstances surrounding the problem.

Suppose the batteries are installed in a radio or some other piece of equipment which is sold, through retailers, to consumers. Then the consequences of a decision may depend on the proportion of batteries that would burn out or otherwise fail in less than 10 hours of use (or whatever other figure management may have found gives rise to consumer complaints). Hence, π would be, in such a situation, the appropriate decision-parameter.

Alternatively, suppose the company uses the batteries internally in its own manufacturing or research operations. If so, it may be interested in how much use the shipment of batteries as a whole might afford. Thus, to think in a per battery figure, it may be interested in the arithmetic mean of the times to failure of the batteries in a shipment. In that case, μ would be the decision-parameter of interest.

In either case, however, the decision-parameter could not be determined by studying the population—if any information was thought desirable, sampling would be necessary. Why?

Suppose a decision-parameter is selected for a problem, yet in analyzing their choice, those in charge of the planning of a study find that even if known, the measure would not afford a basis for decision. What is wrong? The study simply has not been properly planned—further planning is still necessary. Perhaps their choice is a poor one. Perhaps other decision-parameters are needed to supplement the original choice. Perhaps, not a decision-parameter but a frequency distribution must be known. The difficulty may be even more basic. Perhaps the wrong population of study was defined, or perhaps the population is inadequate. The situation can only be remedied by the combined efforts of the subject matter experts and the statistician involved in the problem.

If anything of this sort is wrong, however, the only time to remedy it satisfactorily is in the planning stage. It is too late to do so once a study has come up with incomplete or inadequate answers.

3.9 You Should Now Know That

A second step in statistically defining a problem is specifying the decision-parameter(s).

A decision-parameter is a property of a population which is pertinent to a decision problem.

In many decision problems a decision-parameter presents a way of statistically describing states of the world.

In some problems two or more decision-parameters are required; in others the entire frequency distribution may be needed.

The choice of the decision-parameter(s) in any problem is governed by the answer to the question: "What do we have to know about the population in order to make a decision?"

Selecting the decision-parameter(s) in a problem requires the combined judgments of management and the statistician.

Every property of a population is called a parameter; only if pertinent to the decision problem is a property called a decision-parameter.

The proportion, the aggregate, and the arithmetic mean are three important and very commonly used decision-parameters.

The aggregate is the sum or the total of a set of observations.

The arithmetic mean is a statistical name for what often is called "the average."

Two other simple and occasionally used decision-parameters are the median and the mode.

The range, the average deviation, and the standard deviation are three measures of the variability in a population.

The standard deviation is the most widely used of these population properties.

The square of the standard deviation is called the variance.

Among the symbolic names we have adopted for decision-parameters are:

π : the proportion
 A : the aggregate
 μ : the arithmetic mean
 σ : the standard deviation
 σ^2 : the variance

The normal distribution or normal curve is a particularly important population model.

3.10 You Should Now Be Able to Solve

1. Almost 40 years ago, one outstanding statistician in the field of educational statistics characterized democracy as a frequency distribution of abilities and inclinations. As he stated:

The problem of a democracy is to utilize the talents, attitudes, and training of the individuals of a distribution for the sake of raising the mean of the self-same distribution to higher levels of capacity, attitude, and training. None will deny that the elevation of the social mean is a

good, while there are cogent social arguments against decreasing the variability of the group.⁹

Explain why most people in a democracy would agree with this statement.

2. A firm may classify a product it produces as defective or non-defective and use the proportion defective as a decision-parameter to describe the efficiency of its operation. On the other hand, it might measure some characteristic of the product, such as its width, and use the arithmetic mean as a decision-parameter. What considerations would enter into its choice?

3. A fraternity brother offers to bet you, with handsome odds, that you cannot walk across Mindowaskin Park's lake in Westfield, New Jersey. Consider the population of depths at all parts of the lake. What property of this population is pertinent to your decision problem on whether or not to accept the bet?

4. Two instructors give the same statistics course at your school. One gives a final grade on the basis of the mean of a student's population of quiz grades; the other grades on the basis of the median of a student's population of quiz grades. If you had a choice of with whom to take the course, which teacher would you select? Why? Assume teaching ability, jokes, difficulty of quizzes, and all other pertinent factors constant between the two.

5. Recent articles on education in the United States have used the arithmetic mean salary of teachers as a criterion to demonstrate that teachers are paid low salaries. Discuss the adequacy of the mean as a decision-parameter in this instance. Would you suggest an alternative criterion?

6. What decision-parameter governs the outcome of a presidential or congressional (or most any) election?

7. In each of the following decision problems state the statistical population of interest and the appropriate decision-parameter or decision-parameters. Explain your answers:

- a) A large women's wear chain receives a shipment of 2,000 dresses. Before distributing the dresses to its various outlets it must determine whether the manufacturer has marked the proper sizes on the dresses.
- b) An automatic machine is set to stamp out bottle caps 1.2 inches in diameter. You are to decide whether the machine is operating prop-

⁹Truman L. Kelly "Elementary Statistics in High-School Mathematics as a Socializing Agency," *School and Society*, Vol. XI (1920), pp. 228-30.

erly or not; that is, whether the setting should be left alone or be adjusted.

- c) You want to determine the value of the white goods (sheets, towels, etc.) inventory in a large department store.

8. You apply for a position with a large corporation. The personnel director tells you the entrance salary and the minimum and maximum salaries employees have earned after 5 years with the company. Would these decision-parameters offer sufficient information *on salaries* for you to make a decision on whether or not to accept the position if it were offered to you? If not, what other decision-parameters would you want to know?

9. Your company considers purchasing a special type of thermostat from a manufacturer. That firm's sales engineer, in discussing the quality of the product with you, points out that their production process is carefully set to produce a thermostat which operates at an average temperature of 64 degrees, and that the standard deviation of this population is only .2 degrees. "Therefore," he says, "about 68 per cent of the thermostats we would deliver to you will operate at between 63.8 and 64.2 degrees, and almost all parts will be within your specifications of 63.4 to 64.6 degrees." What assumptions does this statement rest upon? Explain each carefully.

3.11 You Will Also Find That

Most of the statistical measures which we have called decision-parameters and which we have studied in this chapter are discussed in every basic statistics textbook in one framework or another. Although there is some variation in the amount of detail with which books treat averages, variability, and the normal curve, you should find the following of interest:

HIRSCH, WERNER Z. *Introduction to Modern Statistics*, pp. 32-47, 49-64. New York: Macmillan Co., 1957.

MANDEL, B. J. *Statistics for Management*, pp. 95-117, 120-41. Baltimore, Maryland: Dangary Publishing Co., 1956.

NETER, JOHN, and WASSERMAN, WILLIAM. *Fundamental Statistics for Business and Economics*, pp. 187-93. New York: Allyn & Bacon, Inc., 1956.

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics: A New Approach*, pp. 211-41, 244-53. Glencoe, Ill.: The Free Press, 1956.

For additional notes on the early history of the arithmetic mean, the standard deviation, and the normal curve, see:

WALKER, HELEN. *Studies in the History of Statistical Method*. Baltimore, Maryland: Williams & Wilkins Co., 1929.

Sample Selection

4.1 Introduction

The first three chapters have laid the groundwork for our study of statistical decisions. They have presented the steps necessary to define many problems statistically. They have considered how one goes about answering such questions as: "What is the problem?" "What observations are pertinent to my problem?" "What properties of my population of observations would I like to know?" In short, the material we have considered emphasizes how to think about one's problem in statistical terms. Usually, this way of thinking simply sets the stage for the making of decisions; it tells you whether or not there are pertinent observations and how you might use them.

Statistical decisions, of course, are made on the basis of observations, but ordinarily not on the basis of a complete set of them—many are based on only a sample. Hence, once a problem is defined in statistical terms there still remain questions on how many observations to obtain and on how to obtain them. Let's begin with some general remarks on the first of these points: How many?

First of all, you must remember that there are three broad answers to this question: (1) no observations, (2) some observations (that is, a sample), and (3) all pertinent observations (that is, the population completely).

You see, the statistical definition of a problem may put the decision problem in such a perspective that an answer is obvious. Then, no observations would be necessary in order to see which alternative is best. Suppose that we further develop this thought with a very specific example.

Example 4.1 A certain part is made in the machine department of an aircraft parts manufacturer. The parts are then shipped to the assembly department (in a nearby plant) in lots of 2,000. Thus, the decision problem is whether or not a given lot should be inspected before the assembling operation.

Furthermore, the problem can be defined statistically. In brief, a part is an elementary unit; a pertinent observation is whether a part is defective or nondefective. The total set of these observations is the population. The decision-parameter of interest is π —the proportion of defectives in a lot.

Now, let's consider consequences. We may assume that if a lot is inspected prior to its shipment, all defectives can be eliminated (although this is a naïve assumption). Inspection costs, however, average about \$50 a lot.

If a lot is not inspected, any defective parts should be picked up by a later 100 per cent check after the assembly operation. The cost of this operation is \$.10 per item. Therefore, to assemble a defective part is to waste 10 cents.

For any lot of 2,000 we can denote the proportion of defectives by π —this is the decision-parameter of interest. Then the number of defectives is 2,000 times π . Thus, the costs, per lot, associated with the alternatives "no inspection" and "100 per cent inspection" are as follows:

Type of Cost	Alternative	
	No Inspection	100% Inspection
Inspection cost.	0	\$50
Cost of assembling defective parts.	$$.10 \times 2,000 \times \pi$	0
Total Cost.	$\$200 \times \pi$	\$50

For example, for various values of π (various states of the world), the total costs are:

Alternative	State of the World (π)								
	.00	.01	.05	.10	.20	.25	.30	.35	.40
No inspection.	\$ 0	\$ 2	\$10	\$20	\$40	\$50	\$60	\$70	\$80
100% inspection.	50	50	50	50	50	50	50	50	50

What do we conclude? Well, for all values of π less than .25, no inspection is the preferable alternative. Only if a lot is more than 25 per cent defective should 100 per cent inspection be considered. Of course, on any given lot of product the company is uncertain as to the

actual value of π . But the company should be almost certain that π is less than .25. If they are not, something is wrong with the production process, and revamping this process, rather than inspection, is the solution to the problem.

In short, the company should not consider the alternative of 100 per cent inspection at all. Therefore, there is no need to make any observations on a given lot of parts in order to choose between no inspection and 100 per cent inspection.

The choice on whether or not to make observations in this last illustration rests with an analysis of explicit cost considerations. Such a direct analysis is not always possible—but the example still illustrates the type of thinking necessary for deciding whether or not to make observations. In short, as a general rule, one need not make observations in a statistical decision problem unless he actually is uncertain about which course of action is the wisest. For example, there was no such uncertainty in Example 4.1. This was pointed up by the statistical definition of the problem.

However, in many, if not most, problems some degree of uncertainty exists about the true state of the world. Ordinarily, you will find that some observations will be of aid in reducing uncertainty. However, this may very well depend on whether you have to conduct a study yourself or whether you may be able to find information pertinent to your problem already collected. Let's consider this point in detail in the next section.

4.2 Primary and Secondary Data

There are two different ways in which pertinent observations may be obtained for a statistical decision problem. First of all, data may be derived from a study conducted for the specific purpose of providing a basis for decision on the particular problem at hand. Information such as this may be referred to as *primary data*, and the statistical study designed to provide them may be referred to as a *primary study*.

At times, however, the information needed in a decision problem may be obtained from data already published or otherwise in existence. Such data may have been gathered for some other purpose than coping with the problem of immediate concern. Nevertheless, they may be useful in one's current problem. Data such as these may be called *secondary data*, and the study by which they were obtained may be called a *secondary study*.

Generally speaking, all data, in a sense, are at one time or another primary data. Someone had to collect them, and they had to be col-

lected for some purpose. The distinction of primary versus secondary therefore depends on the particular decision problem at hand, and on whether or not the data were collected for that problem.

For example, decisions on whether to accept or to reject a lot of product ordinarily are made by the firm receiving the product on the basis of a sample. Such a study, therefore, is primary—at least from the point of view of the “consumer.” However, suppose a certain lot is returned to the “producer” together with a report of the sample results and the producer uses those results in deciding whether to scrap, rework, or reship the lot. From his point of view and for his decision problem, the data provided by the study are secondary.

In any statistical study we must consider the potential usefulness of both primary data and secondary data as a basis for action. However, in the case of primary data one main question is: “How can we conduct a study to obtain pertinent observations?” In the case of secondary data the corresponding question of interest is: “Were these data obtained by a reliable procedure?” One always should conduct a primary study in accord with certain basic principles of statistics—principles you will learn in the course of our studies. Similarly, one always should evaluate a secondary study by these same basic principles. Data collected for another purpose may or may not be useful for a given decision problem. They may or may not be reliable. They may or may not come up to your standards. We must learn how to judge whether or not they do.

One obvious advantage of using secondary data, if they are reliable, rather than primary data is that ordinarily the former are less expensive to obtain. In fact, many business organizations often *must* rely on secondary data because a primary study may be far too expensive. Remember, we always must weigh the cost of securing numerical information against the benefits (the reduction in uncertainty) to be derived from the use of such data.

Some secondary data are *internal*; that is, part of the records of the firm that arise in carrying out its day-to-day operations of purchasing, producing, selling, shipping, advertising, and so forth. These data may be compiled primarily to set up the accounting records of the firm, yet they also may be drawn upon when needed to provide numerical data for decision problems that face the business.

Since internal records may serve as secondary data for decision, it is essential in planning what records to keep and in what form to keep them that management be aware of the types of problems these data could help to solve. While it is true that many records

must be kept primarily for legal reasons, a little thought about their potential decision-making uses may avoid the need for many special studies. It also may avoid continuous revisions that otherwise would be needed to devise accounting systems to include pertinent numerical data.

Example 4.2 Losses due to rework and scrap generally accompany each and every operation in the production of a product. Obviously, such losses bring about a reduction in profits, and it is important to control them.

One aircraft company has developed an interesting program in its attempts to reduce rework and scrap and thereby cut losses. For one thing, attention is given to those areas where losses are greatest. To determine these areas the company uses sample data to estimate the ratio of rework and scrap losses to the cost of production labor. Such data are available as part of the accounting records of the firm.

To determine the numerator of this ratio (rework and scrap losses) the company's records must provide information as a basis for the allocation of the costs of rework time and scrap to each operation in the production of airplanes. To determine the denominator of the ratio (production labor), records must be kept of the hours (or cost) of direct labor for the same operations. Without such internal data the program to control losses would not be possible. Such records exist at the company because of close co-operation between the accounting and statistical divisions (another instance of the importance of the principle of teamwork).

To make decisions in our highly interdependent economy, management often requires data *external* to its own enterprises. Thus, it may need information on the state of the national economy, on the entire industry, and other macroeconomic data. In recent decades a vast quantity of these data has been gathered. In this connection, the federal government is the largest producer of secondary data in the world. And others, including trade associations, private research organizations, financial services and journals, the United Nations, and state and local governments, contribute pertinent facts in many reports, periodicals, yearbooks, and other sources.

External secondary data usually are provided by some central agency when the potential users of such data are not capable of doing the study themselves. Thus, the federal government almost continuously provides a variety of information about the economy. This is, in many cases, in line with its responsibility of insuring economic stability in the country. But businessmen also make use of these published data on prices or inventories or sales. In part, the

government does the job because it is needed by individuals who could not do it themselves.

An organization that conducts such studies must realize that its job is to put the consumer of these data in a position to act as though the consumer himself has made the study. In cognizance of this, the federal data-collection agencies often consult with local business groups and local government to determine the type of data, and the form in which these data should be, in order that they be useful as bases for action. Similarly, trade associations often set up special committees to study the type of data needed by their members.

It is essential that management be aware of the types of external secondary data that are available. The use of such data avoids the making of decisions in the face of an excessive amount of uncertainty. In addition, it saves the firm the necessary expense of obtaining information already in existence through a study of its own.

Example 4.3 Many residential areas in Nassau County, New York, are serviced only by electricity; no gas service is available. Nevertheless, residents in these areas regularly are swamped by mail advertisements, telephone calls, and door-to-door promotional campaigns to purchase gas refrigerators, dryers, and heating equipment.

This needless expenditure of time and money could be avoided by a request to the Long Island Lighting Company, the supplier of gas and electricity to the area, for a map of the areas that do have gas facilities. Some companies apparently have been unaware that secondary information on the areas serviced by gas is available.

It would be futile to attempt to list and describe all available sources of secondary data here. The study of data sources is a part of many different fields of knowledge. Experts in each substantive field such as economics, marketing, management, and accountancy are acquainted with the sources of data relevant to their own particular field. However, some illustrations should point up the usefulness of secondary data in statistical decision problems.

Example 4.4 The Disston Company is a manufacturer of industrial products such as metalworking items, machine knives, and woodworking items. The company determines the sales potential of its products in various geographical areas. It then uses this information as a guide to evaluate the effectiveness of a distributor in selling the company's product in his area, and acts accordingly.

To determine sales potential, the company relies on secondary data. Members of the research department have stated: "As a multi-line manufacturer of industrial goods . . . we cannot afford to spend

the time in thorough and exhaustive market analysis for each of our products individually; and as a result, the existence of detailed information such as the *Census of Manufactures* is of enormous benefit to us."¹

The *Census of Manufactures* is a survey of all manufacturing activities in the United States. It is conducted under the auspices of the U.S. Bureau of the Census. Published results include extensive detail on employment, payrolls, production, and consumption of raw materials for many industries.

Thus, the Disston Company determines its area potential for veneer knives by consulting the *Census of Manufactures* to find out the quantity of veneer produced by industries who could use their knives: veneer plywood, cigar boxes, fruit and vegetable baskets, and wooden boxes. The demand for knives to cut veneer is then related to the quantity of veneer produced. External secondary data provide a basis for decision.

Example 4.5 According to one authority,² the criteria important to arbitrators in making their decisions about wage awards include the following: (1) changes in the cost of living and real wages, (2) budget studies, (3) wage comparisons, (4) changes in productivity, (5) ability to pay, and (6) the economic environment. Although decisions do not rest solely on numerical information, such data, to be useful, must be timely and reliable.

At times, primary studies may be conducted to provide information for arbitrators on some of these criteria. Often, however, reliable secondary data are available on many of these points.

For example, wage data are compiled by the Bureau of Labor Statistics and published in the *Monthly Labor Review* and in special publications. The same organization publishes comprehensive indexes showing changes in consumer prices. The National Industrial Conference Board, a private research organization, also compiles wage data and publishes a consumer price index. In addition, many trade publications (such as *Editor and Publisher*), trade associations, and state and local governments publish wage data.

Data on worker's budgets are prepared by the U.S. Department of Labor and are available for 134 cities. One study was made in 1947, but the budget can be brought up to date by applying changes in the consumers' price index and in taxes.

Information on productivity is compiled by the Bureau of Labor Statistics of the U. S. Department of Labor for a number of important industries.

Data on industry and total corporate profits (as one measure of

¹ Henry Schweitzer and Philip Torray, "Uses and Development of Industrial Potentials" in Harry Brenner (ed.), *Marketing Research Pays Off* (Pleasantville, N.Y.: Printers' Ink Books, 1955), p. 305.

² Jules Backman, *Economic Data Utilized in Wage Arbitration*, Labor Relations Series (Philadelphia: University of Pennsylvania Press, 1952).

ability to pay) are compiled by: (1) the U.S. Department of Commerce and reported in the *Survey of Current Business*; (2) the Federal Reserve Board and reported in the *Federal Reserve Bulletin*; (3) the Bureau of Internal Revenue and reported in *Statistics of Income*; (4) the Federal Trade Commission and Securities and Exchange Commission and published in quarterly reports; (5) the National City Bank and reported in its *Monthly Letter*. Data on individual companies can be obtained from Moody's and Standard and Poor's financial reports and from newspaper reports.

Current information on the general status of the economy is found in such sources as the *Survey of Current Business* and the *Federal Reserve Bulletin*. Historical series covering many phases of economic life may be found in the *Statistical Abstract of the United States*, a publication of the Department of Commerce.

Many other sources of secondary data also could be listed in connection with this decision problem. Knowledge of the availability of such information and the proper manner of relating observations to decisions in wage arbitration are an integral part of that branch of applied economics.

Example 4.6³ The Federal Trade Commission once issued a complaint against a flour company to the effect that a planned acquisition by the company would substantially lessen competition.

In order to make a proper economic evaluation of the complaint it was necessary to determine what proportion of the total market for flour was controlled by the companies to be merged. Flour is of three types: family, bakery, and mix.

To determine the necessary information the company used the following sources:

1. The total market for all types of flour was obtained from the Department of Agriculture's yearly estimates of total civilian wheat flour consumption.
2. *Family* flour consumption data were provided by estimates from the U.S. Department of Commerce.
3. *Bakery* flour consumption was provided by the *Census of Manufactures*.
4. Trends in brand share of the family and mix flour markets were obtained from studies conducted by the Market Research Corporation of America.
5. Family and mix flour consumer habit patterns for rural and urban areas were obtained from studies conducted for other purposes by the company.
6. Company deliveries of family and mix flour were derived from the company's internal records.
7. Population estimates were obtained from the Bureau of the Census of the U.S. Department of Commerce.

³ This illustration is based on G. R. Detlefsen, "A Procedure for Market Estimating," in H. Brenner (ed.), *op. cit.*, pp. 57-67.

8. National income data were obtained from the U.S. Department of Commerce.

Thus, all the basic data were secondary in nature. In addition, all the data with the exception of those concerning the company's deliveries and the figures from the *Census of Manufactures* were derived from sample studies, and had to be evaluated as such.

The principle of planning must be remembered in secondary studies as well as in primary studies. It is too often forgotten because secondary data usually are less expensive to obtain than primary data. Despite this fact, planning—relating decisions to potential observations—still is necessary. For one thing, the cost of gathering even secondary data is far from trivial. Furthermore, masses of data may be built up—data which confuse rather than clarify the problems facing a decision maker unless planning is implemented.

When we use secondary studies, care must be taken to verify that the data refer to the population of interest. If not, it must be decided whether or not the study provides an adequate working population. Other questions also must be raised: "Are the data sufficient to allow us to determine the decision-parameter(s) of interest?" "Can we have confidence in the data found in secondary studies?"

The auspices under which a study is conducted may be a general indicator of the reliability of secondary data. However, more precise criteria are needed to evaluate the quality of such data. Thus, as noted, whether a study is primary or secondary, we should conduct or evaluate it by certain basic statistical principles—those we are in the process of learning.

4.3 Reasons for Sampling

Whether data are primary or secondary, it is important to recognize that they may be either complete or sample information. In this book our emphasis is upon sampling since statistics is the subject which deals with making decisions in the absence of complete or population data. Thus, when the decision-parameter and, hence, the true state of the world in a business problem is unknown, certain statistical methods and theory may be used to provide a basis for decision. These techniques are based on the theory of sampling. Sampling is important in primary studies for reasons considered below. It also is important when we are dealing with secondary data; if such data are based on a sample study, they must be evaluated carefully.

When a statistician talks about a sample, he means a part of a sta-

tistical population. We have already discussed the latter concept in considerable detail; you should recall that a statistical population is a complete set of pertinent observations, either in the form of variates or attributes.

A *sample*, therefore, simply is some of the pertinent observations from a specified population.

Furthermore, *sampling* is the process by which the sample of observations is obtained from a population. Thus, sampling involves at least two steps: selecting elementary units and making observations on them.

Remember, then, that strictly speaking, a sample refers to a group of observations. Thus, when a statistician "samples," he "samples" observations. However, at times, it is convenient and customary also to speak of a sample of elementary units because such units must be selected before we observe them. Therefore, as is often done, we may use the terms "sample" and "sampling" in two ways: to refer to a well-selected set of observations and to refer to a well-selected set of elementary units.

In either event, however, sample information represents incomplete information. Samples are parts of statistical populations which represent sets of complete information. But why and when should we ever utilize samples? Why settle for less than the whole population? There are a variety of reasons which arise individually or collectively in a large number of business problems. Probably the five most basic and most widely occurring reasons can be termed and listed as follows:

- a) Economy
- b) Timeliness
- c) Destructive nature of a test
- d) Accuracy
- e) Infinite population

Before elaborating and illustrating these notions it should be stressed that not *all* of these reasons are equally important in every business problem. However, often at least one of them must be considered. Also remember that other good reasons for sampling besides these may occur in certain problems. Now for explanations and examples of the five terms given above.

Economy refers to the simple fact that obtaining sample information is often (but not always) less expensive than obtaining complete information.

Example 4.7 A soap company wishes to determine the distribution of lengths of time a standard box of detergent lasts a housewife. The purpose of the study is to decide whether or not to market various sizes of boxes of the detergent. A survey which would include all housewives in the United States obviously would be far too expensive. However, a sample of housewives would provide a cheaper, yet reliable, basis for making the decision.

Timeliness refers to the idea that obtaining sample information generally will result in a considerable saving in time spent on a statistical study and will allow prompt and timely results. In many business problems this time element is a consideration of paramount importance.

Example 4.8 As noted in an earlier section, the federal government and many private business organizations periodically publish numerical information on the United States economy. These statistical data include employment, price, inventory, production, and purchasing figures for selected industries and for the whole country. Unless the information can be brought out rapidly, the time period to which the data would refer makes them of little use as an aid in determining the level of the current economic scene or as a guide in short-run economic forecasting. Sampling, however, does allow for the timely collection of these data.

In order to make observations in some business problems, particularly those dealing with manufactured products, the elementary unit being observed must be destroyed or weakened. To test all units of the population would result in their destruction. It is obvious that sampling is a better alternative.

Example 4.9 A firm produces radio tubes which it would like to sell with a guarantee of 100 hours. How can it estimate the number of tubes that would fail in less time and be returned? Testing all of the radio tubes until they failed is out of the question for there would be none left to sell. However, selecting and testing some of the tubes provides one basis for making a decision.

The fourth point, *accuracy*, is a subtle one. It refers to the fact that very often sample information is more accurate than complete information and, therefore, permits more valid decisions. This holds true whenever a sample study can be conducted more effectively than a complete count.

For example, a statistical survey which involves personally contacting individuals to get information—perhaps market research information—necessitates a group of competent and well-trained interviewers. A complete count ordinarily necessitates a large group

of interviewers; a sample necessitates only a small group. This smaller group is usually easier to train; the larger may suffer from the effects of the law of diminishing returns. Thus, a sample survey with a small, compact, well-trained set of interviewers usually will be able to get more valid information from the public. The result: more accurate results and more valid decisions.

The last argument for sampling, an *infinite population*, should be a fairly obvious point. Generally, it refers to the fact that when we are dealing with an infinite population, as in an analytical study, we *must* settle for a sample from the population.

Example 4.10 A manufacturing process produces coils. Suppose the production manager of the company must decide on a setting for a machine. Since his decision is about the process, the population of study should be defined as the totality of observations (e.g., defective versus nondefective) on all coils that the process could turn out under a given set of operating conditions. Thus, note that any statistical study of the process would be analytical and that the population is infinite.

It follows that a complete count is impossible. Observations on any one or two or three day's product would represent only partial information about the population. If any information is to be the basis for decision, it must be sample information.

In concluding our remarks on reasons for sampling it should be remembered that in some problems a statistician will not recommend the use of sample data. Rather he might say: "Make your decision without the benefit of observations (as in Example 4.1)—it is not feasible to attempt to get any information."

Or, in the other extreme type of problem, he might say, "Let's utilize a complete count and not a sample for this problem—all things considered, it will be more economical." Why should he say such a thing? First of all, the use of samples in decision making always involves some risks—risks of making an incorrect decision because complete information was not available. Thus, a complete count is preferable if no risks can be tolerated.

Example 4.11 As you undoubtedly know, the number of Congressmen from each state in the House of Representatives is based on the population of the states. This is required by the Constitution (Article I, Section 2), and provision is there made for a decennial census of the population in order to so apportion representatives in Congress among the states. Thus, the Constitution provides that an enumeration of the population be made within 3 years after the first meeting of the Congress of the United States and within every subsequent term of 10 years.

Politically, it would not be feasible to base this apportionment decision on sample data—no risk of an incorrect decision due to incomplete information would be tolerated.

Secondly, a complete count is preferable to a sample if a population is small. This is so because the difference between the cost of a sample and the cost of a complete count in such a situation is also small.

Notwithstanding the fact that some decision problems can be solved either without the benefit of observations or with complete information, the large majority are best solved through the use of sample data. This is true when we use primary data, and it is equally true when we use secondary data. The next main questions in our studies, then, are: "How should sample data be obtained?" and "What is a well-selected sample?"

4.4 All Possible Samples

We now should be clear on the idea that sample information is partial information and that this partial information is obtained from a group of only some (not all) elementary units. However, it is important to recognize that in any given problem there are *many different samples* of elementary units that *might* be selected for study. And it is equally important to recognize that in most problems *only one sample* actually is selected.

Before discussing procedures for selecting that one sample, however, let's consider all possible groups of elementary units. Suppose, for example, that you are interested in testing a group of 6 machines—call them A, B, C, D, E, and F. How many different samples of 2 are possible? One way of answering this question is through an enumeration of all possibilities. Thus, we find the following combinations:

AB	BC	CE
AC	BD	CF
AD	BE	DE
AE	BF	DF
AF	CD	EF

These 15 combinations exhaust the set of all possible samples of 2 machines. However, in any test of a sample of 2 machines, one and only one of these combinations actually would be selected.

In general, suppose that we have a set of N elementary units representing a population. How many possible sample of n units are there? For example, we know that the answer to this question for

$N = 6$ and $n = 2$ is 15, and we might be able to answer it for other specific values of N and n on the basis of similar enumerations. However, the general answer can be provided only by a branch of mathematics called combinatorial analysis.

Thus, to express this answer in terms of a formula we must introduce a type of number called a *factorial*. For example, "5 factorial" is written $5!$ and is equal to $5 \times 4 \times 3 \times 2 \times 1 = 120$. Other examples of factorials are:

$$\begin{aligned} 1! &= 1 \\ 2! &= 2 \times 1 = 2 \\ 3! &= 3 \times 2 \times 1 = 6 \\ 4! &= 4 \times 3 \times 2 \times 1 = 24 \\ 10! &= 10 \times 9 \times \dots \times 2 \times 1 = 3,628,800. \end{aligned}$$

As you can see, factorials increase in value very rapidly. If you need further proof consider the facts that

$$20! = 20 \times 19 \times \dots \times 2 \times 1 = 2,432,902,008,176,640,000$$

and $100!$ is almost equal to the number 1 with 158 zeros after it!

We need not consider the values of factorials for anything other than positive integers and zero. Zero factorial, incidentally, is written $0!$ and is, by convention (and for convenience) taken equal to 1. (Thus, $0! = 1!$, since both equal 1.)

The number of possible samples of n elementary units from a population of N elementary units may be expressed in terms of factorials. It equals

$$\frac{N!}{n!(N-n)!}.$$

For example, with $N = 6$ and $n = 2$, we have

$$\frac{6!}{2!4!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 4 \times 3 \times 2 \times 1} = \frac{720}{48} = 15.$$

Similarly, as you may verify:

N	n	Number of Possible Samples
6	3	20
10	4	210
20	6	38,760
100	3	161,700

That is, there are 20 possible samples of 3 from a population of 6, etc.

One last detail—suppose, as in an analytical study, a population

is infinite; that is, N is infinite. Then, there are an infinite number of possible samples of a given size.

Now, with the idea of all possible samples in mind, in the next section let's return to the problem of how we actually select one of these possibilities in a practical problem.

4.5 Probability, Judgment, and Convenience Sampling

The key question to be answered in this section is: "What procedure should be used to select *one* of the *many* possible samples of elementary units that are available in a given study?"

There are many different ways of selecting a sample. It is convenient, however, to divide these many ways into three categories: probability sampling, judgment sampling, and convenience sampling. See if you can attach these suggestive terms to the methods of sample selection which are given in the following illustration.

Example 4.12 An association of purchasing agents, located in New York, periodically conducts a sample survey of its members to determine their buying policies and their expectations of changes in the price level. It reports the results of the study to all members, some of whom use the information as a basis for modifying present purchasing policy.

There are, of course, many ways in which this type of survey might be carried out. Among them are included the following:

- a) The officials of the association might select 200 of its members, including those whom *they considered* "typical" or "representative" of the entire membership. This sample of 200 members then would be asked to provide the desired information, which would be considered "representative" of the population (which involves all members).
- b) Information might be requested only from those members who worked in the New York area because they could be contacted quickly and most easily. Again this information would be a sample from the population of interest.
- c) The officials of the association might write each member's name on a separate slip of paper, place all slips in a hat, mix them thoroughly, and select, say, 200 at random. The members so selected would be contacted to provide the sample information.

If you haven't guessed, method (a) in this example is an illustration of *judgment sampling* (or as it is also called, *purposive sampling*). This method of sample selection calls for an expert in the subject matter—e.g., purchasing or economics—to decide on which elementary units should be included in a study. Thus, for all judgment samples, the judgment of an individual determines the elementary units that should be included in a sample. Usually this judg-

ment is exercised by attempting to include what the expert considers "typical" or "representative" units.

Method (b) represents *convenience sampling*. Expert judgment is not used to decide on the elementary units to be included in a convenience sample—convenience is the only criterion for selection. Any sample derived in this way is, therefore, only a convenient slice of a population—this slice sometimes is called a *chunk* to distinguish it from a well-selected sample. Other examples of chunks are represented by (1) interviews of people sitting in a park because they happen to be there, (2) the selection of one's class of students because they are handy, and (3) comments by customers who complained about some product during the past month because their names are readily available.

In contrast to the use of judgment, or the criterion of convenience, in the selection of a sample is *probability sampling* (or *random sampling*). For the moment we may view a probability sample as one which is selected by some random process. A crude and imperfect example of a probability sample is illustrated by (c) in Example 4.12. More precise and reliable methods of selecting probability samples will be considered later. In any event, probability sampling involves giving every elementary unit some known chance of appearing in a sample.

Throughout this course we will center our attention on probability sampling, for probability sampling is of paramount importance in the field of statistics. Statisticians regard probability sampling as the best method of sample selection because it is the only method of sampling which can be objectively evaluated. (As we will soon see, this evaluation is accomplished through probability theory.) Thus, their train of thought runs along the following lines. How do we use samples? We use them to make decisions when we are uncertain about the value of a decision-parameter. Since there are many possible samples, the one actually selected may be good or bad, and we rarely can determine which it is. If we can, there ordinarily is no need to sample. Thus, a sample may or may not lead to a correct decision—there are always certain risks of making incorrect decisions whenever courses of action are taken on the basis of samples rather than on the basis of complete information. Only the use of probability sampling permits the measurement of these risks.

It still must be remembered, however, that judgment sampling and convenience sampling have more than a few legitimate uses. For one thing, judgment sampling often is preferable to probability sam-

pling if only a small sample is to be used as a basis for decision. How large is small? This depends on the problem situation. For example, if eight or ten counties were to be selected from all counties in the United States for a study of the condition of county roads, it probably would be best to select the samples by using (an expert's engineering) judgment. Experience has led statisticians to expect a smaller risk of error with judgment sampling than with probability sampling in such a case.

Example 4.13 Suppose that a legislative committee conducts an investigation of state taxation of interstate commerce. They undoubtedly would want to check and to study the state tax laws of *all* states as a basis for deciding on what type of federal legislation, if any, to recommend to Congress. However, if a stop-gap law is needed quickly, a sample of one or two or three states' tax laws might be selected for immediate study. If so, it would be best to select those few states on the basis of a tax expert's judgment as to which laws were representative of all the state statutes.

However, judgment sampling often is dangerous, and, unfortunately, it is used far more frequently than it should be in far too many types of decision problems.

Example 4.14 One inadequacy of judgment sampling is pointed up by the results of a sampling experiment conducted by an English statistician, Frank Yates.⁴ A group of about 1,200 stones, varying in size, was spread out on a table, and 12 persons were each asked to select 3 samples of 20 stones, which were to represent, as nearly as possible, the size distribution of the whole collection.

The results of the experiment showed that there was a consistent tendency, common to nearly all persons, to select a sample which overestimated the average size of the stones. In fact, only 6 of the 36 estimates were smaller than the true average weight, and 3 of these estimates were made by a single observer.

Another interesting aspect of such a judgment sampling procedure is the consistent tendency to underestimate the variability in the distribution. Thus, most of the persons selected stones as near their concept of the average size as possible, with a much smaller proportion of extreme sizes than in the whole population.

It must be remembered that judgment is not unimportant in statistics—we already have used it in several ways, and we will use it in many others. For example, judgment is quite necessary in defining a problem, in defining a statistical population, and in specifying the decision-parameter. However, except in rare cases, judgment is un-

⁴ See Frank Yates, *Sampling Methods for Censuses and Surveys* (London: Charles Griffin & Co., Ltd., 1949), p. 12.

necessary and even undesirable as a basis for deciding which elementary units should be selected. Hence, except where noted, selection by probability sampling is to be preferred to judgment selection.

Convenience sampling, too, often is misused. Convenience sampling may lead to reliable results, or it may not. As with judgment sampling, there is no way of evaluating it. Convenience sampling often is particularly dangerous when one assumes the sample so selected will be "representative" of the population.

Example 4.15 One once overly popular method of sample selection in marketing research is called *quota sampling*. A quota sample is, in everyday language, supposedly, a "perfect cross section" of the population. Unfortunately, there are several defects in quota sampling, and many experiences have shown it a complete failure.

In the design of a quota sample, certain quotas are set—for example, for the number of males and females to be included in the sample, for the number of urban and rural families to be included, for the number of young, middle-aged, and older persons to be included, etc. Thus, the sample is distributed over various groupings so that it will agree almost perfectly with known data in regard to such criteria as area, sex, age groups, economic levels, etc.

However, it is important to recognize that this will *not* guarantee a "good" sample. In fact, this type of sample may fail completely to correspond to the population of interest *with respect to the characteristics that the study is supposed to measure* (even though it corresponds in several characteristics already known and used to construct quotas).

Once quotas are set, they then often are given to interviewers to fill as best they can—by searching out people with the appropriate characteristics. To this extent, quota sampling is "convenient." Thus, despite its pretense at careful planning, the actual selection of elementary units is left to interviewers and their convenience.

Example 4.16 The experience of the *Literary Digest* (a once-popular but now defunct magazine) in predicting the outcome of the 1936 presidential election is cited widely in statistical circles as an example of the danger of interpreting any sample as reliable.

The previous political predictions of the *Digest* poll had been good at times. For example, in 1932 it predicted the vote for Franklin D. Roosevelt within 1 percentage point. However, in 1936 it predicted the defeat of Roosevelt and the election of his Republican opponent, Alfred M. Landon. Actually Roosevelt won, receiving 60 per cent of the votes cast. The *Digest* poll had failed to pick the winner—what was worse the poll made an error of 19 percentage points in predicting Roosevelt's vote.

The *Digest's* predictions were based on a convenience sample—and on a rather large one at that. From a mailing list of about 10

million names, more than 2 million ballots were received and included in the poll. The *Digest* settled for the returns received—they were conveniently available. This illustrates the point that even a large convenience sample may yield poor results. Increased sample size does not cover up the inadequacies of a poorly selected sample.

Thus, statisticians regard convenience sampling as useful only when one needs rough estimates, or when a probability or judgment sample is impossible, or when an important decision does not hinge on the data to be collected. In addition, convenience sampling is particularly useful when the purpose of a study is only to produce ideas rather than to make decisions.

Example 4.17 A baby food products manufacturer continuously engages in research to bring the country's infants better puddings, cereals, strained fruits, etc. In this connection, parents are contacted and they and their children are requested to try certain foods. Parents are asked to comment on the foods. Of course, not all parents (nor children) are willing to try these new products and provide comments—hence, those that do represent only a convenience sample.

Fortunately, then, important marketing decisions are not made on the basis of this type of information. However, since the purpose of such a study merely is to develop ideas on color, appearance, flavor, etc., of baby foods, a convenience sample suffices.

Example 4.18 Convenience samples are used extensively in *motivation research*. Such studies as applied to marketing generally attempt to determine why people buy. They involve the use of such psychological techniques as depth interviews, word association, narrative projection, cartoons, and personality tests. Motivation research stresses the importance of individual and social needs as determinants of buying power in addition to economic and physical needs.

Motivation research very often is used for exploratory purposes—to give ideas to management as to which alternative courses of action are available. In such instances convenience samples are very effective. However, when the function becomes more than that of exploration of possible alternative courses of action, the selection of respondents and sample size should be determined on the basis of probability sampling theory.⁵

Recently a motivation research study to determine why people prefer particular makes of automobiles was conducted to consider possible contents of automobile advertisements. The study indicated the possible conclusion that the purchase of a car hinges more on

⁵ Joseph W. Newman, *Motivation Research and Marketing Management* (Boston: Harvard University, Graduate School of Business Administration, Division of Research, 1955).

suiting the personality of the car to the personality of the buyer than on the engineering features of the car. The report thus focused attention on alternative forms of advertising to project the personality of automobiles.

In short, then, probability sampling—selection by means of some random process—enables one to measure risks of incorrect decisions objectively. Since statisticians consider this of paramount importance, a good part of our study of statistics will center on it. Judgment and convenience sampling will be dealt with only insofar as we need contrast them to probability sampling. However, for this reason you should keep them in mind.

Also remember that no matter whether a sample is selected by judgment, by convenience, or by a random process, the principle of planning must still be applied. Careful consideration of how the sample information will be used is invariably necessary to prevent the collection of unnecessary data.

You may have noticed that so far we have not considered how one should decide on the size of a sample. Often the number of elementary units to be selected for study is governed by the amount of time and money management is willing to spend to provide a basis for decision. More objective ways of determining sample size also are available when we use probability sampling, but these depend on theory we have yet to develop and must be discussed later.

Let's now turn to a fuller discussion of probability sampling.

4.6 Simple Random Sampling

Probability sampling, as was noted in the last section, is a method of obtaining partial information. It involves selecting elementary units by a random process. When probability sampling is used, we can determine the chance that a given elementary unit will be selected for study and this, in turn, will enable us to determine risks of incorrect decisions.

It must be noted that there are many different types of probability samples, just as there are many different ways of exercising judgment or applying the criterion of convenience in those sampling procedures. Thus, in the remaining pages of this chapter we will consider several different methods of probability sampling and how we can apply them in administratively efficient ways. As a start in this direction, suppose that we consider the simplest, most basic, and most common type of all probability sampling procedures—*simple random sampling*.

Simple random sampling may be defined as a method of selection which gives *every* possible sample of size n the same chance of being *the* sample chosen in a study.

For example, suppose we have a population represented by five college professors and including their subjects of specialization: economics, psychology, sociology, history, and government. All possible samples of two include the following (in terms of first letters):

EP	EH	PS	PG	SG
ES	EG	PH	SH	HG

If we use a procedure of selection which gives each of these 10 combinations an equal chance to be *the* one sample selected, we are using a method of sampling called simple random sampling.

How, in a practical problem, can a simple random sample be selected? Are there any ways of carrying out the procedure we have defined? There are—in fact there are a variety of ways. In the next section we will turn to a more sophisticated method, but for the present consider a procedure similar to one suggested earlier: Identify every possible sample combination (AB, AC, AD, etc.) on a slip of paper, throw all the slips in a hat, and select one of them at random. Such a procedure, if carefully applied, should give every possible *combination* of elementary units the same chance of being the one sample realized in a given problem. Hence, it should meet the definition of simple random sampling given above.

However, one refinement in our procedure for selecting a simple random sample is desirable even when we sample by drawing slips of paper from a hat. Rather than choosing combinations of elementary units, we may choose units one at a time. Thus, to select a simple random sample it may be shown that all we need do is use a method of selection which gives every elementary unit the same chance of being the *first* unit selected, every *remaining* elementary unit the same chance of being the *second* unit selected, every *remaining* elementary unit the same chance of being the *third* unit selected, and so on.

Ordinarily it is easier to select a simple random sample by picking elementary units one at a time. For one thing, this allows us to bypass the problem of enumerating all possible samples, the number of which is usually very large.

Therefore, it is important to recognize that under this alternative procedure, it may be shown that every possible sample has a known (and equal) chance of being the one sample actually selected. In

the next chapter we will express these known chances in numerical terms—in terms of a probability measure. This, in turn, will allow us to calculate risks of making incorrect decisions on the basis of sample information.

For the present simply note that this idea of “calculable risk” is *the* advantage of any method of probability sampling. This is important because there are several popular misconceptions about probability sampling. For example, probability sampling does not guarantee a “fair,” or a “typical,” or a “representative” sample, though some people loosely and incorrectly describe probability sampling by these terms. For one thing, such terms are vague and usually meaningless. For another, no method of *sampling* can guarantee fair or typical or representative *samples*. Generally speaking, there is no way of evaluating samples. However, with probability sampling we can evaluate the sampling procedure in terms of the risks it involves. This idea of “calculable risk” is its important advantage.

The misconception of fairness is particularly prevalent in the case of simple random sampling and stems from the fact that simple random sampling is thought of as a procedure which gives every elementary unit the same chance of being selected. First of all, it must be recognized that many other methods of probability sampling also give every elementary unit the same chance of appearing in a sample. For example, with a population of six elementary units, A, B, C, D, E, and F, we might pick one of the three combinations, AB, CD, or EF at random. You can see this procedure gives every elementary unit an equal chance. However, it does not give every combination of elementary units the same chance. (Several including AC, AD, AE, etc., would have no chance.) Thus, the procedure is a probability sample of some sort, but it is not a simple random sample.

Secondly, note that some methods of probability sampling, by design, may give some elementary units a better chance of being selected than others. For example, in a sample study of profits of motion-picture theaters one might wish to give Grauman's Chinese Theater and Radio City Music Hall a larger chance of appearing in a sample than Neighborhood Theater or Hometown Cinema. The important characteristic of probability sampling is the fact that every elementary unit has an ascertainable chance of being selected—these chances may be equal or unequal.

In our discussion of probability sampling, we implicitly have limited ourselves to finite populations, i.e., enumerative studies. Now,

the matter of probability sampling from infinite populations (for analytical studies) deserves some attention.

One of our standard references for analytical studies has been a decision problem dealing with a production process. Since in this type of problem a manufacturer is interested in taking action on a process, the population of study is infinite. It is represented by the totality of product the process could produce if it continued to operate indefinitely under a given set of conditions. Now, we have defined probability sampling as the random selection of part of a population so that every elementary unit has a known chance of appearing in the sample. What does this mean for an analytical study?

First of all, you must realize that in sampling from such an infinite population we are limited to selecting from one or more days' output. That is, even though we want to select from the total population (all possible product the process could produce), we operationally are limited to a part of this total population (what has been produced today or over the last few days). How do we know, then, that we have a probability sample from the total population when we are not controlling the selection procedure?

The answer to this question depends on the nature of the production process. Experience may have taught us that such a process itself operates in a random manner. Thus, we may assume a production process generates a probability sample of all product it could produce.

Thus, even if we were to take a complete count of all product produced on a given day, we would get a probability sample from the total population. At least we would if our assumption that the process is generating some type of probability sample is a good assumption. Furthermore, if we were to select part of any day's output, we also would wind up with a probability sample of the total population. Again, however, this result depends upon our assumption that the process is operating randomly.

These assumptions may be tested, in some cases, by statistical methods on the basis of our experience with the process under study. That is, for a given type of probability sample our sample should behave in a certain way, and empirically we can see whether or not it is behaving in that manner. For example, in flipping a coin, observing H T H T H T H T H T H T would make us suspicious that our flipping process was not generating a simple random sample—each observation seems to depend on the previous one.

4.7 Random Number Tables

The only method we have considered for drawing a probability sample thus far has involved slips of paper and a hat. However, this procedure should not be used in practice because it presents several difficulties. First of all, the slips of paper representing the various combinations of elementary units or the elementary units themselves should be almost exactly the same size and weight. They cannot be too heterogeneous; otherwise their respective chances of being selected would vary significantly. Furthermore, the mixing of the slips before one is selected must be more thorough than most people can imagine. Unless everything is ideal in every sense of the word, we would not be using a procedure which met our definition of simple random sampling. What's worse, we might end up thinking that we were selecting a simple random sample and our thinking would be all wrong.

To top everything else, writing each possible sample or even writing elementary units on separate slips of paper in order to select a sample is a rather cumbersome and costly task—especially when the total number of elementary units is rather large. Then, too, a rather large hat would be needed for most practical problems. Thus, a more economical procedure is absolutely necessary.

Fortunately, statisticians have developed such a procedure. It involves the use of something called *random numbers*. In this connection the following quotation is pertinent:

It was recognized by some workers early in the development of statistics that many of the methods which may reasonably be expected to yield random samples are, in fact, biassed. The evidence which has since accumulated supports that view. It seems that wherever any human element of choice is allowed free play, as, for instance, when an observer selects "at random" by eye a number of plants in a field, or draws cards "haphazardly" from a pack, bias inevitably creeps in. There may be individuals whose psychological processes are so finely balanced that they can deliberately choose random samples; but they are the exception, not the rule, and the ordinary statistician dare not use his own nervous mechanism as a random selector.

There is thus a need for an objective method of drawing random samples, which form the basis of a large section of theoretical and practical statistical studies. In the search for a reliable method many devices have been tried, but those which offer the greatest *a priori* prospect of randomness are unfortunately difficult to wield in practice . . . The method of Random Numbers is the most reliable and often the speediest method of drawing random samples . . . Other methods may be quicker and just

as adequate for particular [problems] . . . but no other method is known which can in practice be used without misgiving on any [problem].⁶

Now, perhaps it is best if first we let some of these random numbers speak for themselves. A few of them are given in Appendix C. This group represents a section from a much larger table, but they will suffice for our present studies. Look them over.

The numbers you have observed are called random because they typify exactly what is meant by the word. They represent a thorough scrambling of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. No matter how you read them—sideways, down, upside down, or backwards—they still seem to have that random property. It seems that every digit is independent of the preceding one, of the succeeding one, of the upper one, and of the lower one. These numbers are typical of what we mean by random phenomena.

There are several tables of random numbers which are used extensively in statistical practice.⁷ Appendix C reproduces a portion of one such table. The digits given come from a very much larger table compiled for use in sampling studies by The Rand Corporation. Tables of random numbers have been produced in a variety of ways. The mathematical and technical details involved cannot be considered here. Needless to say, one way of *not* constructing such a table is to write down digits as they come into your mind. Such a tabulation undoubtedly would be biased in several ways—for example, by a preponderance of your favorite digits or your lucky number.

How can we determine whether the numbers in a table of random digits are actually in a random order? One answer is that the random nature of such a table of digits can be determined empirically. First of all, there are various statistical tests that can be applied to see if all the digits occur about the same number of times, and that 2's do not have a habit of following 6's, etc. Secondly, there is always the pragmatic test—do these numbers serve the purpose we wish them to? Well, to answer this let's see how we might use them.

In general, we wish to use random numbers to help us in selecting probability samples. Suppose, then, we first consider how we would use them to select a simple random sample. Variations of this

⁶ M. G. Kendall and B. Babington Smith, "Tables of Random Sampling Numbers," *Tracts for Computers No. XXIV* (London: Cambridge University Press, 1939), p. vii.

⁷ The earliest published table (1927) was compiled by L. H. C. Tippett. From British Census Reports, Tippett selected 40,000 digits at random and combined them into 10,000 groups of 4 digits each.

procedure also serve for many other types of probability sampling.

The first thing we must do in using random numbers is to consider the *frame*, that important ingredient of many statistical studies that we considered in Chapter 2. A frame is a working definition of a population. It identifies all of the elementary units representing the population of study in terms of a list or map or in some other concrete way.

Given a frame, we must assign a unique serial number to each and every elementary unit. This identifies the elementary units numerically. For example, if we were to sample from a list of 10 warehouses (A, B, . . . , I, J) we could assign A the serial number 1, B the serial number 2, . . . , I the serial number 9, and J the serial number 0.

One thing of importance to remember in assigning serial numbers is that they all should be different. Another is that they should all include the same number of digits. Thus, suppose we had a frame listing 678 warehouses. We would have to assign 678 different serial numbers to the elementary units. Since the largest serial number to be assigned is a 3-digit number, each serial number may be taken as a 3-digit number. Thus, we may assign the first warehouse the number 001 (not plain 1 since it must have 3 digits), the second 002, . . . , the thirty-third 033, . . . , the one hundred and twelfth 112, . . . , the last, 678.

Once serial numbers are assigned to the elementary units included in the frame, we may use random numbers to select the elementary units to be included in a sample. For example, to select 5 warehouses from the 678 noted above, we turn to our table of random numbers and read off 3-digit numbers. In practice, one usually will pick up where he left off the last time he used the table. However, since this is our first experience we may start at the beginning and, for convenience, read digits in groups of 3 down the table. Refer to Appendix C and observe that this means we read off 931, 282, 218, 220, etc.

These numbers identify certain serial numbers and, hence, certain warehouses. Thus, the warehouses corresponding to these random numbers are selected for our sample—those selected are observed for the characteristic of interest. How many warehouses should be selected? The first 5 warehouses we can identify. Why 5? Because we want a sample of 5. In general, if we want a simple random sample of size n , we continue using our table of random numbers until we identify n different elementary units.

However, as you may have asked: "What about that 931 above?" Since we had only 678 warehouses, no warehouse was assigned that serial number. We may skip any set of digits that has no elementary unit associated with it. Furthermore, if any serial number comes up twice, we skip it the second time—we ordinarily should not want to include any elementary unit more than once in a sample.

Although we have used our random number table from the top, since the numbers are in a completely random order we might have started anywhere at all in the table. However, as was noted, in practice one usually should pick up where he left off; he makes note of where he finishes using the table for one sample and proceeds from there on his next sampling problem.

Thus, the use of tables of random numbers to select a simple random sample (or some other type of probability sample) generally involves two distinct steps: (1) assigning serial numbers to every unit in the frame, and (2) selecting a sample of these units by reference to the table of random digits. In various problems, of course, the ways in which these steps may be taken vary according to the circumstances of the situation.

Example 4.19 In defining and classifying jobs, pertinent information includes the proportion of time spent by present employees in various activities. One statistical application which arises in this connection is called work sampling. Work sampling involves the selection of times, at random, and observing the activities of the employee under study at the times selected.

The sample of times may be selected by the use of random numbers. For example, suppose an employee's activities from 8:00 A.M. to noon and 1:00 P.M. to 5:00 P.M. are to be sampled on a given day. First, each minute of time may be assigned a 3-digit serial number, perhaps as follows (note that the list of times provides a frame):

<i>Time</i>	<i>Serial Number</i>
8:00-8:59.....	800-859
9:00-9:59.....	900-959
10:00-10:59.....	000-059
11:00-11:59.....	700-759
1:00-1:59.....	100-159
2:00-2:59.....	200-259
3:00-3:59.....	300-359
4:00-4:59.....	400-459

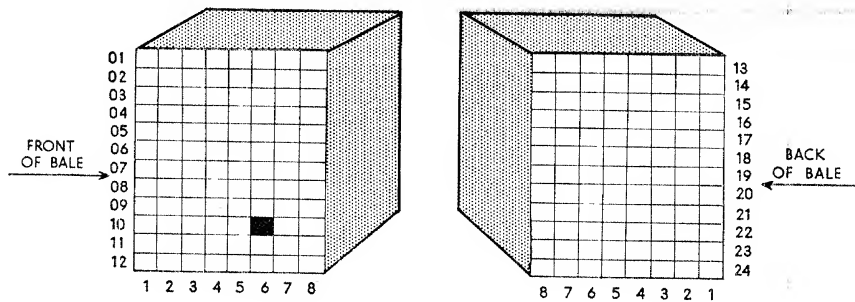
Note also that the use of 000-059 for 10:00-10:59 and 700-759 for 11:00-11:59 avoids 4-digit serial numbers and other 3-digit serial numbers already used.

Random numbers then would be used to identify the times to be selected. For example, using a table of random numbers we may find the digits 413, 952, 886, etc. The 413 identifies 4:13 P.M. which is one

time for the sample; the 952 identifies 9:52 A.M. which is a second time for the sample; the 886 identifies nothing and is skipped; etc.

Example 4.20 As was pointed out in an example in an earlier chapter, wool ordinarily is imported in highly compressed bales. Suppose then, that an importer, facing the decision problem of how to grade a bale, wishes to sample the wool to provide a basis for his choice of an alternative. How might we select such a sample?

One solution is to subdivide (on paper) the bale into "cores" about 4 inches in diameter and half the depth of the bale in length. Serial numbers then should be attached to each core. For example, the 2 sides of the bale might be drawn as follows:



Note that this diagram represents a frame and that each "core" is identified by a pair of serial numbers—e.g., (6,10) is marked on the diagram above. A sample of "cores" would be selected by reading out pairs of random numbers from a table. Those points that fall in the sample would then be cored out and tested as a basis for the classification decision.

In many accounting and auditing applications by contrast with the situations just described, the problem of assigning serial numbers is no problem at all. Many elementary units, such as checks, tickets, invoices, vouchers, etc., are numbered as part of regular accounting operations. When these prenumbered units represent a population of interest, the first task in selecting a probability sample is already solved, and the selection of the sample is greatly simplified.

Is the use of random numbers in situations such as those we have covered in this section a good procedure for selecting probability samples? This question can be answered empirically as well as abstractly. If the repeated use of such tables shows us that the chances of serial numbers appearing are, for all intents and purposes, equal, then we may conclude that the use of random numbers is justified. Experience has taught statisticians that this, in fact, is true. Hence, the use of tables of random numbers, as we have sketched it, brings us very close to insuring the theoretical definition of simple random

sampling (or some other type of probability sampling) in practice. Furthermore, it usually is a rather simple way of achieving this goal.

4.8 Other Types of Probability Samples

In this text we shall concentrate on the use of simple random sampling as a means of obtaining partial information. One reason for this is its relative simplicity. Another is that this type of probability sampling is basic to an understanding of almost all other sampling procedures—many other methods merely are variants or combinations of simple random sampling. Lastly, simple random sampling itself is an important and frequently used sampling procedure.

However, to point up that there are other types of probability samples, we will devote this section to a brief outline of four other methods: (a) stratified sampling, (b) cluster sampling, (c) double sampling, and (d) sequential sampling. Most of the details of the theory and the exact applications of these methods are out of the range of this book. Thus, we should regard this survey only as a quick look at what the above terms refer to.

All of the aforementioned methods of sampling are methods of probability sampling. That is, despite their differences, each procedure is designed to give each and every elementary unit a known chance of appearing in a sample. Therefore, all should be drawn by using a randomizing mechanism such as a table of random numbers.

Of course, all of these methods of sampling differ in one or more important respects. In any given problem, they may differ in their cost per unit sampled. Thus, some probability sampling methods are more expensive to use than others. Then, too, some methods may provide a more reliable procedure than others for some specific problem—this means less risk.

One statistical problem in the design of a sample study is to take account of these differences in cost and risk, and to determine and use the most economical probability sampling procedure available. Suppose now we briefly consider the potential usefulness of each of these four methods as compared to simple random sampling.

Stratified sampling involves partitioning or stratifying a population into groups of elementary units and selecting a simple random sample from each group or stratum.

A stratified sample is, obviously, not a simple random sample. Note, for example, that with four elementary units, A, B, C, and D, simple random sampling gives every possible sample—AB, AC, AD, BC, BD, and CD—the same chance of being selected. However, sup-

pose that we divided the population into two strata: (A, B) and (C, D). If one unit were selected from each stratum, the only samples with a chance of occurring are AC, AD, BC, and BD. Neither AB nor CD are possible with stratification.

Among the principle reasons for stratification and, hence, for stratified sampling are (1) information for the various strata of the population, as well as the population as a whole, is needed; (2) stratified sampling may be more convenient than simple random sampling, as in the case of a survey being carried out over a wide geographical area where different field offices could each supervise the sample study of each stratum; and (3) the use of stratification may reduce the risks of relying on sample information.

Example 4.21 Suppose that your college wished to determine the extent to which various factors influenced students' decisions to attend that particular school. The college might need such information in deciding on how to slant its promotional and public relations program and on how to modify its admissions policy.

In selecting a sample of students from whom to obtain information, a stratified sampling procedure might be wise. For example, one basis for stratification would be freshman, sophomore, junior, and senior. Thus, if a total sample of 200 were to be selected, 50 might be picked from each class *at random*. Alternately, the population might be stratified by the criterion of whether or not the students were from outside of, or within, the college's home state. Then, perhaps a simple random sample of 150 in-state and a simple random sample of 50 out-of-state students would be chosen. Either or both of these criteria might be a useful basis for stratification. If both were used, 8 strata would be formed and simple random samples selected from each.

The advantage of stratification in such a problem might arise from the need for information about each stratum as well as about the population as a whole. Or, if not, it might arise because stratified sampling would involve a smaller risk than simple random sampling.

Note, for example, that if the class-year basis of stratification were adopted, a sample of *all* freshman would be impossible (only 50 could be freshmen). However, with simple random sampling such a sample has a small chance of occurring. Note that this would be disadvantageous only if freshmen had different reasons for attending the school than other students. In such a case the sample would be unrepresentative in the *characteristic of interest*. The risk of this possibility is eliminated with stratification.

Generally speaking, whenever strata differ with respect to the characteristic of interest, stratified sampling involves less risk than simple random sampling.

Stratified sampling often is confused with quota sampling (described in Example 4.15) because both involve subdividing the pop-

ulation into different groups. While it is true that they are similar, they differ in one important respect. The elementary units to be covered by a stratified sample are drawn at random; those covered by a quota sample are drawn by convenience. Furthermore, it is important to remember that neither method guarantees a "good" sample—both involve risks. However, under stratified sampling these risks may be measured in terms of probability, while under quota sampling there is no way of evaluating them.

In addition, it may be noted that useful stratification, whatever its purpose, demands some a priori knowledge of the population to provide a basis for stratification. Generally speaking the problem of providing this basis is a nonstatistical one.

Cluster sampling involves partitioning a population into groups or clusters of elementary units and selecting a simple random sample of clusters.

For example, suppose we want to select a cluster sample of 40 dwelling units from a population of 1,000. We might form 250 clusters of 4 neighboring homes in the population. A simple random sample of 10 clusters would give us a total sample of 40 dwelling units.

You can see the main advantage of cluster sampling from this illustration. It is represented by the savings in time and money that accrue from the fact that some elementary units falling into our sample will be close together (those in the same cluster). However, this also points up a disadvantage in cluster sampling. Roughly speaking, neighboring elementary units may have similar characteristics. Hence, any cluster ordinarily is more or less homogeneous—it cannot be expected to provide as much different information about a population as, for example, four independently selected elementary units.

The methods of selecting probability samples that we have considered up to now all have at least one characteristic in common—they are all fixed-size sampling methods. That is, the sample size is chosen a priori, and this number is selected. After, and only after, the whole sample is drawn are observations made and the appropriate action taken.

However, some methods of probability sampling do not share this characteristic. For example, *double sampling* involves plans for selecting one's sample in two stages. One sample is drawn, and depending on what is observed, some action is taken *or* a second sample is selected. Of course, if a second sample is selected, action then *must* be taken.

Ordinarily a double sampling plan will require a smaller total sample than simple random sampling to achieve the same reduction in uncertainty. This is its main advantage. Its relative complexity is one administrative disadvantage.

Sequential sampling is an extension of the basic idea of double sampling. However, elementary units are selected and observed one by one and a decision on whether to take a certain action or continue sampling made after each observation.

In conclusion, it must be pointed out that no matter what type of sampling procedure is to be used—simple random, sequential, or cluster, etc.—one must recognize that risks are involved. Different probability sampling designs lead to different risks, but in all cases these risks may be measured in terms of probability (our next chapter's topic).

4.9 You Should Now Know That

After defining a decision problem statistically, one must determine whether any observations actually are necessary.

If they are, information may be provided by a primary study—one conducted to provide a basis for decision on the particular problem at hand.

Some pertinent information may have been collected for some other purpose, yet it may be useful in one's immediate problem. Such information is called secondary data.

Secondary data may be internal or external, depending on the auspices under which they were collected.

Pertinent information may be derived on the basis of a sample or a complete count.

A sample is part of a population. Sampling is the process by which a sample is obtained.

Five possible reasons for sample selection rather than a complete count of a population are (1) economy, (2) timeliness, (3) destructive nature of a test, (4) accuracy, and (5) an infinite population.

Decision making on the basis of sample information is risky. These risks can be evaluated objectively only if probability sampling is used.

Probability sampling is a selection procedure in which every elementary unit has an ascertainable chance of appearing in the sample.

Judgment samples are those for which elementary units are selected on the basis of an expert's knowledge of the field of inquiry.

Convenience samples are those for which elementary units are selected because they are readily available.

One important type of probability sample is called a simple random sample. It is one which is drawn in such a manner that every possible sample of a given size has an equal chance of being selected.

Probability samples are drawn rather easily through the proper use of random numbers.

Other probability sampling procedures include stratified sampling, cluster sampling, double sampling, and sequential sampling.

4.10 You Should Now Be Able to Solve

1. A retailer has data, ranging back 2 years, on the number of sales that he has made daily. He also has data on the total dollar sales per day. Are these primary or secondary data? Do the data represent complete or sample information? Explain your answers.

2. In many automobiles gears are shifted by means of push buttons. Assume that an automobile manufacturer receives a shipment of push-button panels to be installed in cars. An incorrectly wired panel would result in a car going in reverse when the drive button is pushed, and vice versa. Of course, if an improperly wired panel is installed, it would be detected when the completed car is driven off the assembly line.

Should the manufacturer check the panels in deciding whether or not to install them?

If so, should he check the entire lot or a sample selected from the lot to decide whether or not to install the panels?

3. In each of the following instances indicate whether you would suggest use of a convenience, judgment, or probability sample. Explain the reasons for your choice:

- a) A manufacturer of an industrial product desires to prepare a questionnaire to explore potential new uses for his product. To help design the questionnaire, a sample of present users and nonusers of the product is to be interviewed.
 - b) A company is to select 100 workers from its 3,000 employees to determine the attitude of its employees towards a recently introduced social activities program. It will use this information to decide on whether the program is operating smoothly and should be continued, or whether changes should be made.
 - c) The school districts of a state have been in difficult financial straits because of the rapid increase in pupil enrollments. The state's department of education, therefore, decides to undertake a detailed study of the problems of five school districts.
-

4. In each of the following instances describe how you would select the required sample. (First determine the population of interest and a suitable frame.)

- a) A simple random sample of 50 from invoices for a given month that are numbered from 2,463 to 3,981 consecutively.
- b) Samples of apples from trees to compare the effectiveness of different sprays.
- c) A simple random sample of entries in an accounts receivable ledger. The ledger contains 150 accounts with varying numbers of debit and credit entries and is to be checked for accuracy.
- d) A sample of sea water at a beach to determine bacteria count before approval of the beach by the Board of Health for public use.

5. Appraise the following methods of sample selection.

- a) The "inquiring reporter" device used by newspapers to determine opinions on public issues.
- b) The letters received by Congressmen as a guide to the attitude of their constituents in reference to a given bill.

6. A company installs, on a trial basis, a new safety (accident prevention) program in its manufacturing plant. The month before the new program there were 15 accidents (per 1,000 workers). The first month of the new program there were only 10 accidents (per 1,000 workers). Should the new program be continued? Why or why not?

7. A time-and-motion-study man clocks a particular polishing operation 10 times. His results are (in minutes): .15, .13, .14, .16, .12, .15, .15, .14, .16, .15. Is this a probability sample? If so, what type of probability sample is it? What population is it from?

8. The Bureau of Labor Statistics asks the prices monthly (or in some cities, quarterly) of certain items of food in connection with the Consumer Price Index.

- a) The food items are chosen on the basis of judgment. Why?
- b) The stores from which price quotations are taken are chosen at random. Why?

4.11 You Will Also Find That

Clear discussions of sources of secondary data are included in the following:

BLAIR, MORRIS MYERS. *Elementary Statistics*, chap. xxiv. Rev. ed. New York: Henry Holt & Co., Inc., 1952.

CROXTON, FREDERICK E., and COWDEN, DUDLEY J. *Applied General Statistics*, pp. 45-49. 2d ed. New York: Prentice-Hall, Inc., 1955.

STOCKTON, JOHN R. *Business Statistics*, chap. iv. Cincinnati, Ohio: South-western Publishing Co., 1958.

Pertinent references on sampling, the reasons for sampling, types of sampling procedures, and how to sample include the following:

MANDEL, B. J. *Statistics for Management*, pp. 150-79. Baltimore, Maryland: Dangary Publishing Co., 1956.

NEISWANGER, WILLIAM A. *Elementary Statistical Methods*, pp. 105-44. Rev. ed. New York: Macmillan Co., 1956.

NETER, JOHN, and WASSERMAN, WILLIAM. *Fundamental Statistics for Business and Economics*, pp. 267-81. New York: Allyn & Bacon, Inc., 1956.

CHAPTER - 5

Probability Theory

5.1 Statistics and Statistics and Statistics

We began our study of statistics by studying the nature of business problems and by considering how to translate some of these problems into statistical terms—i.e., by defining the population and by selecting the decision-parameter(s). We then considered the fact that many decisions must be made on the basis of sample studies since complete information is not always available, or may be too costly or time consuming to obtain. Such studies involve risks. Therefore, a statistician ordinarily insists on a probability sample as a basis for decision because only then can these risks be evaluated objectively. In the last chapter we learned how to select probability samples.

So now we have business problems, on the one hand, and knowledge of how to select probability samples, on the other. Our next task is to shake hands with ourselves; that is, to co-ordinate these two aspects of our subject matter. Perhaps the basic questions in which we are interested can be summed up by saying: "We know how to draw probability samples—but what can we do with them? How can they help us in making decisions—how can they help us in solving managerial problems?"

First of all we must realize that in sample studies, despite the risks involved, decisions must be made on the basis of information contained in the sample. Ordinarily, however, a sample will contain some information which is not relevant to making a decision. For example, when a sample of screws is taken from a lot and their diameters measured, we ordinarily are not interested in the array of these diameters. We are interested in the sample average and/or the sample variance, or the proportion falling outside of specifications

—but only for what they indicate about the pertinent characteristics of population. For instance, the *sample* average is important for what it can tell us about the *population* average.

Thus, the sample information which is relevant to the decision parameter(s) of interest is what concerns us, and this information usually can be summed up in such measures as the sample proportion, or the sample mean, or the sample standard deviation. These measures are called *statistics*. Thus, we have a third use for this term. Previously, we have spoken about the fact that statistics are data, and that statistics is the field we are studying. Now we shall speak of a statistic as a measure that summarizes information from a sample.

You might have noticed that the aforementioned statistics have names similar to some decision-parameters. These statistics and the corresponding decision-parameters are almost identical insofar as their definitions or formulas are concerned. The main difference between the two types of measures—and it is an important one—is that a decision-parameter summarizes information about a population, while a statistic only summarizes information given in a sample.

Since we will be concerned with sample observations and statistics, it will be convenient to represent them by some generalized notation. We will continue to denote the size of a sample by n and the size of a population by N . In addition, the various individual observations in our sample will be represented by

$$x_1, x_2, x_3, \dots, x_n.$$

The x_1 represents the first observation in our sample, x_2 the second observation in our sample, etc. The x_n is, of course, the last observation. In general, x_i is the i th observation in our sample.

Notice that in this notation we are using (small letter) x as a symbol for an observation. Previously, in Chapter 3, we used (capital) X in a similar connection. However, at that time we were talking about observations in a population (X_1 was the first observation in the population, etc.). These values were constants—properties of the elementary units under consideration. Now the story is entirely different. The observations that fall into a sample are variables, and to emphasize this difference we use small letters. The value of any x can be any one of the N population values, depending upon which one of all the possible samples is selected and upon the order in which the items are selected. Furthermore, since we are to select probability samples, the sample observations are obtained by means of a

random process and, therefore, they are called *random variables*. The use of x thus connotes the fact that each value appearing in our sample is a result of chance selection.

We used Greek letters as symbols for summary population measures—decision-parameters. We now will use Latin letters as symbols for summary sample measures—statistics. Thus, the sample mean will be denoted by $\bar{x}$ (read x -bar) and expressed in a formula as

$$\bar{x} = \frac{\sum x_i}{n} \quad (i = 1, 2, \dots, n).$$

You should remember that Σ (capital *sigma*) is the summation sign with start and stop instructions stated at the right. Thus,

$$\sum x_i \quad (i = 1, 2, \dots, n).$$

means $x_1 + x_2 + \dots + x_n$; namely, the sum of all n variates which fall into our sample. Thus, $\bar{x}$, the sample mean, is defined as this sum divided by n .

The sample variance will be denoted by s^2 with the formula

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} \quad (i = 1, 2, \dots, n).$$

Furthermore, the sample standard deviation (the square root of the sample variance) will be denoted by s with the formula

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \quad (i = 1, 2, \dots, n).$$

The sample mean and sample standard deviation are measures derived from samples where the observations are variates. When the observations are attributes, the sample proportion often is the statistic of interest. The sample proportion will be called p . It is computed in the same manner as π , though p is a statistic while π is a decision-parameter. For example, suppose that a shipment of resistors were received by a manufacturer. If a simple random sample of 25 resistors were selected and tested, and 2 were found defective, the proportion defective in the sample would be

$$p = \frac{2}{25} = .08.$$

The definitions of the sample measures $\bar{x}$, s^2 or s , and p are similar to those of their counterparts in the population, μ , σ^2 or σ , and π respectively. It is important to recognize, however, that the similarity is only in form since μ , σ^2 or σ , and π always refer to a given

population. They are properties of that population and have only one value—they are invariant. Their values remain the same whether we know them or not. They are decision-parameters, and in no way depend on the sample we select.

However, $\bar{x}$, s^2 or s , and p always refer to a sample. They are summary measures of the information in the sample. If we were to draw samples of the same size over and over again from the same population, $\bar{x}$, s^2 or s , and p would vary from sample to sample. Their values depend on which sample of all possible samples is selected. Statistics, therefore, are variables.

Nevertheless, since in many problems we are uncertain about the exact value of a decision-parameter, we may be forced to rely on the corresponding statistic as a criterion for decision. This is risky because statistics are variables. However, when they are computed from probability samples, they, like the individual observations, are random variables—and the way in which they vary can be predicted. Let's consider the implications of the use of a statistic instead of a decision-parameter as a basis for decision by way of an illustration.

Example 5.1 A soap company must decide what proportion of its current output should be devoted to a more expensive detergent than its present product. If it knew π , the proportion of its customers who would buy the new product, it could make its decision. For example, suppose the firm plans to produce immediately 10,000 boxes of detergent. If it knew π , it would produce $10,000\pi$ boxes of the new product and $10,000(1 - \pi)$ boxes of the old detergent. (Note that $1 - \pi$ is, by definition, the proportion of customers in the population who say they would buy the old detergent.)

In the absence of an economical procedure for obtaining complete knowledge about π , the company might draw a sample of customers. A reasonable statistic summarizing pertinent information about this sample would be p , the sample proportion of customers who would buy the new detergent.

One statistical decision rule for the company to use in its problem, then, would be to produce $10,000p$ boxes of the new product and $10,000(1 - p)$ boxes of the old product. That is, in taking action the firm might use p in place of π , although in doing so they would be taking some risk.

There are many possible values of p that could be obtained in a situation such as that in the preceding example. The actual value of p obtained depends upon which one of all possible samples is selected. Hence, there are risks involved in taking action based upon such a statistic. However, if there were some way of determining the pattern of variability of p , we would have an indication of the pos-

sible risk involved in utilizing the sample proportion (a statistic) rather than the population proportion (a decision-parameter) as our criterion for decision. Thus, roughly speaking, if we determined that 95 per cent of all possible sample proportions ranged from 40 per cent to 60 per cent, we would know that the risk is greater than if 95 per cent of the possible sample proportions ranged from 48 to 52 per cent. In general, we know that the greater the variability in the possible values that a statistic can attain, the greater the risk involved in the use of the statistic as a criterion for decision.

Fortunately, the mathematical theory of probability enables us to determine the pattern of variability of random variables. But statistics are random variables since they summarize observations that are obtained by a random process. Therefore, probability theory will enable us to determine the pattern of the variability of statistics and, hence, to measure sampling risks.

In the remainder of this chapter, we shall deal with the elements of probability theory and see how this theory can be helpful in measuring the risks involved in basing decisions on sample information.

5.2 Mathematical Probability

Probability, chance, and likely are three words which are quite common in everyday conversation. We have all made, or at least heard, statements such as: the probability of a head when a coin is flipped is one half; the likelihood that there is life on Mars is low; the chances that the Yankees or the White Sox will win the pennant next year are very high; and the probability of holding four aces in a game of poker is less than the probability of holding a royal flush.

In which of these connections or in what other situations have you expressed similar ideas? What did you mean by them? It will be advantageous for you to give a few minutes of thought to these questions. Later you can check your intuitive notions against the more formal development given below. But for the present, your intuition should help you to appreciate some of our introductory remarks.

Our present aim is not to consider the term "probability" and its synonyms in their everyday context. Rather it is to give these ideas a technical, rigorous, and specific meaning. Thus, succeeding sections present some of the basic principles of *mathematical probability* as they have been laid down by mathematicians and statisticians over the past two or three centuries. This type of probability may not conform exactly to your intuitive feeling for the term as it is used in everyday conversation, but it should be somewhat similar. In any

event, our short survey will involve, initially, the principles of the theory of probability as a branch of mathematics. Secondly, it will involve the applications of the results of mathematical probability in the field of statistics.

Incidentally, the difference between mathematics and statistics is borne out in the last two sentences. In short, the development of the theory of probability is mathematics, while one of the applications of this theory is statistics. The distinction may be clearer if we draw an analogy with geometry. This subject, as you may know, is developed mathematically; that is, its development hinges on mathematical principles: reasoning from axioms and postulates. The task of applying the results of this development to problems in the real world then falls to the engineer, the architect, the astronomer, or others in applied fields. In exactly the same way it is the statistician who may apply the mathematically derived results of probability theory. Nevertheless, it must be remembered that the foundation of statistics and statistical applications is the mathematical theory of probability.

As we have mentioned, mathematicians and statisticians like to give a more precise and definite meaning to probability than do most people in everyday conversation. What, then, is their definition of probability? This is an easier question to ask than to answer in an elementary textbook. We might hint at the difficulty by saying that there are many ideas on how probability should be defined. There is, however, one point of almost universal agreement. Almost everyone in the field of mathematics and statistics agrees that mathematical probability should be expressed quantitatively. Thus, their attention is centered upon a numerical measure of probability, or as it is sometimes called, a *probability measure*.

The setting for our problem is paralleled by a very familiar phenomenon, the weather. In everyday conversation, we often use the words "hot," "cold," "warm," "cool," "pleasant," etc. These have only a very general qualitative meaning. On the other hand, we are able to measure temperature on either a centigrade or a Fahrenheit thermometer. We are able to give temperature a quantitative meaning. This is precisely what we want to do with probability—to measure it on a scale. Thus, we shall interpret our probability measure as follows:

The probability of an outcome, A , is defined as the proportion of times A would occur in an infinite series of repeated trials of the same kind.

If we let A be some specified outcome in a random experiment, the symbol $P(A)$ will refer to the probability that A will occur. Then our definition may be expressed in a formula as

$$P(A) = \frac{\text{NUMBER OF TIMES } A \text{ OCCURS}}{\text{NUMBER OF REPEATED TRIALS OF THE SAME KIND}}$$

Thus, when we talk about the probability of an outcome, we mean the relative frequency with which it would occur if the experiment were repeated an infinite number of times. For example, we defined simple random sampling as a method of selection which gives every possible sample of size n the same chance of being chosen. What does this mean? It means that to select such a sample we must use a method which if repeated indefinitely would result in each combination appearing with the same relative frequency.

There are two aspects of the definition of probability that we must discuss: first, the notion of "infinite series" and "long run"; and, secondly, the idea of "repeated trials of the same kind."

The idea of "infinite series" and "long run" may be cleared up by a conventional illustration. In terms of our definition, what would we mean if we said that the probability of a coin falling heads when it is flipped is equal to .50? We mean that if this same process, flipping the coin, were continued indefinitely under the same conditions, we would expect 50 per cent of the outcomes to be heads.

We should also point out what the .50 does not mean. It does not necessarily imply that as we continue tossing the coin, we will come to a stage when the number of heads will always be equal to the number of tails. In fact, the difference between the number of heads and number of tails may very well get larger as we increase the number of flips. Thus, in 100 tosses of the coin we might obtain 60 heads and 40 tails, or an excess of 20 heads. After 1,000,000 tosses, we might have 500,500 heads and 499,500 tails, or an excess of 1,000 heads. The relative frequency of heads, however, has decreased from .60 to .5005—very close to our seemingly "good" estimate that the probability of a head is .50.

How do we know that the probability of a head is .50 when we flip a coin? The answer to this question is that we don't know—50 per cent is only a conjecture based on the apparent symmetry of the coin. We would know the true probability exactly only after observing an unlimited number of flips of the coin. This obviously is impossible. A little thought, in fact, will lead you to the conclusion that we can never determine any probability exactly. To do so would

require an infinity of observations! What can be done so that we can ease ourselves out of this quandary? Fortunately, even though we can never know any probabilities exactly, there are various procedures for estimating them.

One method of estimating probabilities is on the basis of past experience. Two examples will illustrate this idea.

Example 5.2 Suppose that an office manager wishes to determine the probability of a paper being misfiled by his file clerks. He examines 1,000 papers filed during the past week and finds 30 misfiled papers. On the assumption that the process of filing or misfiling papers is a random process, an estimate of the probability of a given paper being misfiled, $P(A)$, is

$$P(A) = \frac{30}{1,000} = .03.$$

This value indicates that in the long run approximately .03 of the papers will be misfiled. Although the office manager cannot predict which individual papers will be misfiled on the basis of this information, the estimate does give him some insight into the efficiency of the filing process.

Example 5.3 In setting their rates, life insurance companies make use of mortality tables which give them the probability of a person dying during any one year of his life. These tables are based on past experience, gained by keeping records on ages at which persons die.

For example, one mortality table (there are several, but all give similar probabilities) lists the probability of a 21-year-old person living to 22 as approximately .992. Since he can only live or die, the probability of his dying before reaching 22 is approximately .008.

Here again the process is viewed as a random process. Probabilities such as these, of course, do not give probabilities strictly applicable to any one individual, who might be rather sick or very, very healthy. The probabilities refer to the population at large.

The second method of estimating probabilities is one based on the nature of the process we are observing. For example, when a coin is flipped, there are two possible outcomes, heads and tails; when a die is thrown, there are six possibilities, 1, 2, 3, 4, 5, and 6. Ordinarily, coins and dice are *symmetrical*. Unless we have reason to believe otherwise, it seems logical to *assume* that each possible outcome is *equally likely* or *equiprobable*. Thus, if a die is thrown, what is the probability a 4 shows? Using our equiprobable assumption, $P(4)$ equals $\frac{1}{6}$ since there is only one way a 4 can occur and six different outcomes are possible.

As an illustration of a statistical application of the equiprobable assumption, reconsider the use of random number tables for selecting probability samples. Because of the nature of these tables and the ways in which they are derived, it is reasonable to assume that every digit, 0, 1, 2, . . . , 9 is equally likely in such a table. That is, we may assume that the probability of any digit occurring is $\frac{1}{10}$. Likewise, we may assume that every pair of digits, 00, 01, 02, . . . , 99, is equally likely—that is, each has a probability of $\frac{1}{100}$ of occurring.

The equiprobable rule for estimating probabilities must be used with care. For example, just because you can pass or fail your course in statistics, you wouldn't care to say that both outcomes are equiprobable.

We now turn to the second idea in our definition—the matter of repeated trials of the same kind. This requires some qualification because we have assumed that trials can be repeated. However, a man can only live or die once during his 21st year; in testing a given light bulb it will or won't burn out in its first 100 hours of use; etc. These experiments cannot be repeated. This presents a problem when we attempt to estimate a probability on the basis of past experience because if a trial cannot be repeated we wind up with very little experience.

The problem is usually resolved by pooling many trials of the *same kind*. To estimate probabilities of living and dying we observe not one individual but many individuals of the same kind (although no two individuals are the same, we group those with the same age, occupation, race, etc.). When we flip a coin repeatedly, we assume that the flips are of the same kind (although they may not be exactly so; for example, the coin wears and/or your finger gets calloused).

No trial is exactly repeatable; no two trials are of exactly the same kind in any business situation. Thus, in determining the probability of a misfiled paper in Example 5.2, the trials might not have been of the same kind. The file clerks might have changed, the lighting system could have been improved, or a new supervisor might have been appointed. All these factors might result in trials not of the same kind. The best we can do is try to lump together those that are approximately of the same kind when we must estimate a probability. But even barring this remedy, our definition of probability still provides us with a theoretical bench mark. Remember, then, that $P(A)$ means the proportion of times A would occur if the ex-

periment *were repeated* an unlimited number of times under the same conditions.

5.3 Properties of a Probability Measure

Now, we turn to discuss three basic properties of a probability measure. These properties follow almost immediately from our definition of probability.

First of all, what are the numerical limits of a probability measure? What are its minimum and maximum values? The answer is that for any outcome the probability measure may have a value as small as 0 or as large as 1.

These extreme values of 0 and 1 have a convenient interpretation. Thus, if an outcome never occurs, no matter how many trials are attempted, the probability is considered 0. This is reasonable in view of our interpretation of probability; if an event never occurs then its relative frequency is 0. Furthermore, if one outcome always results from an actual experiment, the probability of this outcome may be considered as equal to 1. This, too, is consistent with the relative frequency definition of probability, as you easily can verify.

For example, on the basis of past experience we might estimate the probability of a baby being born with green hair as 0. On the other hand, we might also say that the probability that every person alive today will die at some time is 1.

A second property of a probability measure is that the sum of the probabilities of all possible outcomes is equal to 1. For example, suppose that we select at random one entry from an accounts payable ledger in order to check its accuracy. The two possible outcomes of this experiment are the selection of an incorrect entry and the selection of a correct entry. Therefore, the probability of an incorrect entry plus the probability of a correct entry must equal 1. Thus, if the probability of selecting a correct entry is .8, the probability of selecting an incorrect entry must be .2, since they are the only possible outcomes. Furthermore, suppose we were to select two entries from the accounts payable ledger. Then 0, 1, or 2 of these may be incorrect. Whatever the probability of each of these outcomes, their sum must be 1.

The third and last property of a probability measure refers to the likelihood that one or the other of two possible outcomes occurs. We will consider this property only as it applies to outcomes that are *mutually exclusive*. Mutually exclusive outcomes are *those which*

cannot occur at the same time. Some illustrations should fix the idea in your mind.

Example 5.4 A personnel administrator must decide on one of several applicants for a position with his company. His alternatives, which include hiring J. Smith, H. Jones, M. Jackson, etc., are mutually exclusive since he is to hire only one of the applicants.

A worker may be absent exactly 0, 1, 2, 3, 4, or 5 days in a given week. All of these possibilities are mutually exclusive. Thus, for example, if he is absent exactly twice, he cannot be absent exactly once.

A simple random sample of n students is to be selected and their average age calculated. Among the possible outcomes of this random experiment are $\bar{x} = 22.1$ years and $\bar{x} = 20.5$ years—two outcomes which are mutually exclusive. If the sample mean is 22.1 years, it cannot be 20.5 years.

A simple random sample of 2 machines from a group of 6 machines (A, B, C, D, E, and F) is selected. If the sample AB occurs, AC cannot occur—nor can AD, AE, etc. All of these outcomes are mutually exclusive.

The rule expressing the probability of mutually exclusive outcomes may be stated as follows: *The probability of either of two mutually exclusive outcomes occurring is equal to the sum of their individual probabilities.* If we write $P(A \text{ or } B)$ for “the probability of either A or B,” we can summarize this rule as

$$P(A \text{ or } B) = P(A) + P(B).$$

This rule is sometimes called the *addition rule*. We must remember that this rule is valid only if the two outcomes are mutually exclusive.

This rule may be extended to three or more mutually exclusive outcomes as well. Thus, if A, B, C, etc., are all mutually exclusive:

$$P(A \text{ or } B \text{ or } C \text{ or } \dots) = P(A) + P(B) + P(C) + \dots$$

Example 5.5 Suppose that the probabilities of exactly 0, 1, 2, 3, 4, etc., machines being out of order at the same time in a plant have been estimated on the basis of past experience. Let $P(0)$ be the probability of 0 machines being out of order, $P(1)$ the probability of 1 machine, etc. Then the information can be summarized as follows:

$$\begin{aligned} P(0) &= .449 \\ P(1) &= .360 \\ P(2) &= .144 \\ P(3) &= .038 \\ P(4) &= .008 \\ P(5 \text{ or more}) &= .001. \end{aligned}$$

What is the probability of exactly 0 or exactly 1 machine being out of order at the same time? Since 0 and 1 are mutually exclusive outcomes, the addition formula can be applied:

$$P(0 \text{ or } 1) = P(0) + P(1) = .449 + .360 = .809.$$

What is the probability of 2 or less machines being out of order? Similar reasoning shows that

$$\begin{aligned} P(0 \text{ or } 1 \text{ or } 2) &= P(0) + P(1) + P(2) \\ &= .449 + .360 + .144 \\ &= .953. \end{aligned}$$

You will recall that the addition rule is *not* applicable when the outcomes are not mutually exclusive. For example, if you were to select a student at random from your statistics class, the probability of selecting a junior or a student over 21 years of age does not refer to mutually exclusive outcomes. A student may be both a junior and over 21 years at the same time; one outcome does not exclude the other. Other simple rules have been developed to cover the calculations of probabilities for nonmutually exclusive outcomes, but we shall not consider them in this text.

5.4 Repeated Trials—Conditional Probability

Up to this point we have dealt with probabilities of outcomes in a single random experiment or trial. For example, we spoke of the probability of selecting a correct entry in an accounts payable ledger, and of the probability of 1 or 2 machines being out of order at one time. In this section, we shall determine how to compute the probability of a given sequence of outcomes in repeated experiments or trials.

Example 5.6 A coin is tossed 3 successive times—these represent 3 repeated trials. One question dealing with these trials that might be of interest is: “What is the probability of the sequence head-head-head?”

An employee is to report to work 5 days in a given week—this situation presents 5 repeated trials. Each trial may result in the worker being absent or present. The type of question we will attempt to answer in this section is exemplified by “What is the probability of the sequence absent-absent-present-present-present?”

Three machines are observed for a week to determine whether they require maintenance work or not. Each machine represents 1 trial—the 3 machines represent 3 trials.

In developing rules for computing probabilities of a given sequence of outcomes from a series of repeated trials, there are two

basic ideas to keep in mind. First of all, we desire rules for expressing the probability of a sequence of outcomes in terms of the probabilities of individual outcomes. This situation is similar to the one we met and solved in Section 5.3 where $P(A \text{ or } B)$ was expressed in terms of $P(A)$ and $P(B)$ —as their sum. Thus, to compute probabilities of sequences we will have to know individual probabilities in order to apply the rules to be developed.

Secondly, it is useful to distinguish between two types of repeated trials and to develop two different rules for these types. Thus, repeated trials may be *dependent* or *independent*, depending upon whether any one trial influences the others or not. For example, two successive flips of a coin should not influence each other—hence, they are independent. On the other hand, suppose that we were to draw two cards from a deck. The outcome of the first draw would influence the outcome of the second (e.g., if the ace of clubs were drawn first it could not be drawn second). Thus, these trials would be dependent.

In this section, we shall develop a rule for computing the probability of a sequence of outcomes when the trials are dependent; later we shall consider independent trials. Thus, the rule for calculating the probability of a specified sequence of two outcomes in two dependent trials states: *The probability of a sequence of two outcomes is equal to the probability of the first outcome multiplied by the probability of the second outcome given that the first outcome has occurred.* You will observe that the probability of the second outcome is conditional upon the occurrence of the first outcome. Hence, it is called a conditional probability. The rule can be stated symbolically as follows:

$$P(A \text{ and } B) = P(A) \times P(B|A).$$

The vertical line in $(B|A)$ stands for the word “given.” Let us look at an illustration of the rule.

Example 5.7 Two forms of a general achievement test are to be selected at random from a group of 6. These tests are to be administered to entering college freshmen. The forms are numbered 1 through 6 inclusive. What is the probability that forms 2 and 4 will be selected *in that order*? When we select the first form of the test, we have 6 tests to draw from. If we use random numbers, the probability of form 2 being drawn first is $\frac{1}{6}$. However, if form 2 is selected, the next test must be drawn from the remaining 5. The probability of drawing form 4 is $\frac{1}{5}$. The drawing of form 4 as the second selection is

conditional upon the drawing of form 2 first. The probability that forms 2 and 4 will be drawn in that order is

$$\begin{aligned} P(2 \text{ AND } 4) &= P(2) \times P(4|2) \\ &= \frac{1}{6} \times \frac{1}{5} \\ &= \frac{1}{30} \end{aligned}$$

This answer may be verified by enumerating the set of all possible sequences of two tests as follows:

1,2	2,1	3,1	4,1	5,1	6,1
1,3	2,3	3,2	4,2	5,2	6,2
1,4	2,4	3,4	4,3	5,3	6,3
1,5	2,5	3,5	4,5	5,4	6,4
1,6	2,6	3,6	4,6	5,6	6,5

When a random method of selection is used, each of these 30 different sequences is equally likely. Form 2 followed by form 4 represents only 1 of the 30 possible outcomes, and, hence, its probability of occurrence is $\frac{1}{30}$ —the same value as that which was obtained by the application of the conditional probability rule.

The conditional probability rule can be generalized for more than two outcomes as follows:

$$P(A \text{ AND } B \text{ AND } C \text{ AND } \dots) = P(A) \times P(B|A) \times P(C|A \text{ AND } B) \times \dots,$$

where $P(C|A \text{ AND } B)$ refers to the probability of C given that A and B have occurred, etc. Thus the probability of a sequence of outcomes in repeated random trials is equal to the product of the probabilities conditional upon the occurrence of the preceding outcomes in the sequence. For example, reconsider the situation described in Example 5.7 and assume 3 test forms are to be selected. What is the probability that forms 2, 4, and 5 would be selected *in that order*? The probability of selecting form 2 first is $\frac{1}{6}$; and the probability of selecting form 4 second, given that form 2 has been selected first, is $\frac{1}{5}$. In addition, the probability of selecting form 5 third, given that forms 2 and 4 have already been drawn, is $\frac{1}{4}$. Hence, the probability of forms 2, 4, and 5 being chosen in that order is

$$P(2 \text{ AND } 4 \text{ AND } 5) = \frac{1}{6} \times \frac{1}{5} \times \frac{1}{4} = \frac{1}{120}.$$

The conditional probability rule for two outcomes and its extension for three or more outcomes are quite useful in the development of statistical theory. In Chapter 4 it was pointed out that to select a simple random sample it is not necessary to enumerate all possible

combinations of elementary units and select one of these at random. An alternative procedure was suggested—randomly select n elementary units one at a time. We have progressed sufficiently in our study of probability theory to indicate the validity of this statement.

Example 5.8 Suppose we consider a population represented by 5 elementary units which may be denoted by A, B, C, D, and E. Any procedure for selecting a simple random sample of 2 must give each possible combination of 2 elementary units

AB	AD	BC	BE	CE
AC	AE	BD	CD	DE

the same chance of being selected. That is, each of these 10 possible samples must have a probability of $\frac{1}{10}$ of being selected.

Suppose, then, we were to select 1 elementary unit from the population at random and follow this by a second random selection from the remaining 4 elementary units. Does this method give every possible sample of 2 the same chance (namely, $\frac{1}{10}$) of being drawn?

Let us answer this question for the sample represented by BD. Thus, consider the probability that units B and D will be selected *in any order*. We must consider the possible order D and then B as well as the order B and then D. The probability of D and then B is, by the conditional probability rule,

$$\frac{1}{5} \times \frac{1}{4} = \frac{1}{20}.$$

But, by the same reasoning, the probability of B and then D also is $\frac{1}{20}$.

These two sequences are mutually exclusive outcomes for they cannot occur at the same time. Thus, the probability of one or the other is, by the addition rule,

$$\frac{1}{20} + \frac{1}{20} = \frac{2}{20} = \frac{1}{10}.$$

This is the probability of the sample B and D occurring in any order. In the same way we may show that the probability of any such sample also is $\frac{1}{10}$.

These results indicate that the method suggested in Chapter 4 for selecting a simple random sample by randomly drawing elementary units one at a time is valid. It is not necessary to enumerate all possible samples and select one of these at random.

You will note that in this last illustration both the conditional probability rule and the mutually exclusive rule were used. Since these rules sometimes are confused with each other, a comment on their applicability is in order. The conditional probability rule is applied to determine the chance that some particular sequence of

outcomes (e.g., B and D) occurs in a series of trials. The mutually exclusive rule is applied in such a situation only to determine the probability of one *or* another of two sequences (BD *or* DB) occurring. Though both rules often are used in the same problem, remember that they are applied to answer different questions.

Just as we used probability theory to indicate the justification of one statistical sampling procedure in Example 5.8, we may use it to point up the validity of other statements made during our study of sample selection. For one thing, it was noted that *several* methods of probability sampling give each elementary unit the same chance of being chosen—that simple random sampling is not the only procedure with this property.

Example 5.9 Assume that we have a population represented by 500 elementary units—for example, 500 index cards in a file. We wish to select a simple random sample of 2 cards. What is the probability that a particular card will be drawn into the sample?

Any one card may appear as the first item selected or as the second item selected. The probability it appears first is $\frac{1}{500}$ because there is a total of 500 cards. Any one card may be the second item selected only if *another* card is selected first. By the conditional probability rule, the probability of this sequence is

$$\frac{499}{500} \times \frac{1}{499} = \frac{1}{500}.$$

Note that the $\frac{499}{500}$ represents the chance of a given unit not being selected on the first draw, while the $\frac{1}{499}$ is the probability of it being selected on the second draw, given that it was not selected on the first draw.

Thus, the probability that a card is drawn first is $\frac{1}{500}$, and the probability that it is drawn second is $\frac{1}{500}$. Since these are mutually exclusive outcomes, the probability that a card will be drawn first or second is

$$\frac{1}{500} + \frac{1}{500} = \frac{2}{500}.$$

You may notice that this is the ratio of the sample size ($n = 2$) to the population size ($N = 500$). In general, it can be shown that with simple random sampling each and every elementary unit has a probability of n/N of appearing in the sample selected.

Example 5.10 Cluster sampling was defined in the last chapter as a method of selection which involves dividing a population into clusters of elementary units and selecting a simple random sample of clusters.

For example, suppose cluster sampling is to be used in a sample study to obtain information on automobile part sales for a state's 2,800 gasoline stations. The 2,800 elementary units (gasoline stations) might be grouped into 5 stations per cluster, resulting in 560 clusters. Then, if a simple random sample of 100 clusters were selected, this would represent 500 gasoline stations.

Since the sample of 100 clusters from the total of 560 is to be a simple random one, the probability of a particular cluster being selected in the sample is $100/560$. In general, the probability of a cluster being selected in a cluster sample is equal to the ratio of the number of clusters in the sample to the total number of clusters.

In addition, each one of the 2,800 gas stations is part of one cluster. Hence, each of their probabilities of being selected is $100/560$. Thus, each and every elementary unit has the same chance of being selected with cluster sampling. This property is not unique with simple random sampling.

You will notice that from Examples 5.9 and 5.10 that both simple random sampling and cluster sampling are methods of selection which give each elementary unit a *known* chance of appearing in a sample and that probability theory enables us to determine these chances. While these chances are all the same in the preceding illustrations, it is important to recognize that for some types of probability sampling these chances at times are unequal. However, equal or unequal, they must be *known* for the sample to be a probability sample.

Example 5.11 As an illustration of a sampling procedure with unequal probabilities, let us reconsider stratified sampling. Stratified sampling, as was noted in Chapter 4, involves partitioning a population into strata of elementary units and selecting a simple random sample from each stratum.

Thus, suppose a firm's management planned to select a sample of 100 employees to determine the extent of their health insurance coverage—the information to be used as a basis for deciding on whether or not to institute a group health insurance plan.

The firm's employees first might be divided into three strata: A—production workers, B—office workers, and C—others. Thus, assume that the sizes of the strata and the samples are as follows:

Stratum	Number of Employees	Sample Size
A.....	600	60
B.....	300	20
C.....	100	20
Total...	1,000	100

Since a simple random sample is to be selected from each stratum, the probability of an employee being selected from a given stratum is equal to the ratio of the sample size to the stratum size. Thus,

$$P(\text{A GIVEN EMPLOYEE IN STRATUM A BEING SELECTED}) = \frac{60}{600} = .100$$

$$P(\text{A GIVEN EMPLOYEE IN STRATUM B BEING SELECTED}) = \frac{20}{300} = .067$$

$$P(\text{A GIVEN EMPLOYEE IN STRATUM C BEING SELECTED}) = \frac{20}{100} = .200.$$

Note that even though the probabilities of an employee being selected vary from stratum to stratum, these probabilities are *known*.

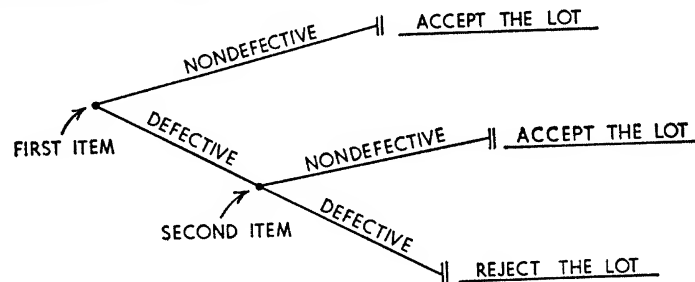
Alternately, suppose that samples of 60, 30, and 10 were drawn from strata A, B, and C, respectively. What would be the probability of an employee being selected from each stratum, in that case?

The probability concepts and rules we have considered enable us to measure the chances of elementary units and combinations of elementary units appearing in a sample. This is important in the theoretical development of statistics. But probability theory does more—it provides us with a scale on which we can measure risks of relying on sampling information.

Example 5.12 A manufacturer receives shipments of a destructible product in lots of 100. He uses small samples to decide whether or not to accept the shipments. One statistical decision rule that he uses is:

Select one item at random. If the item is nondefective, accept the lot. If the item is defective, select a second. If this is nondefective, accept the lot. If the second also is defective, reject the lot.

Schematically, we may view this decision rule as follows:



Since this decision rule is based on a sample, there are risks in applying it. These risks may be measured in terms of probability. However, it is important to recognize that they depend on which state of the world is true. That is, if we measure the quality of a lot by π , the risks of accepting or rejecting the lot depend on the true but

unknown value of π . Because its value is unknown, we consider various possible values of π in order to evaluate the decision rule.

For example, suppose a lot of 100 contains 1 defective—that is, $\pi = .01$. The above rule cannot lead to such a lot being rejected. To reject any lot we must observe 2 consecutive defectives—with only 1 in a lot this is impossible. Hence, the probability of rejecting any lot with $\pi = .01$ is zero—it cannot occur.

Suppose we consider another possible state of the world, $\pi = .50$ —that is, 50 of 100 items in the lot are defective. What are the chances of selecting 2 defectives in a row from such a lot and, hence, of rejecting it? Using the conditional probability rule we find

$$\begin{aligned} P(\text{REJECTING LOT WITH } \pi = .50) &= \\ P(\text{DEFECTIVE, DEFECTIVE}) &= \frac{50}{100} \times \frac{49}{99} \\ &= \frac{2,450}{9,900} \\ &= .248. \end{aligned}$$

The probability of rejecting any lot with $\pi = .50$ is only .248. The probability of accepting such a lot is $1 - .248 = .752$. This indicates that we run the risk of accepting lots with as many as half of the items defective approximately three quarters of the time.

Thus, the risks involved in relying on the suggested decision rule are very large. Nevertheless, such rules have been and are being used in industry. The application of some simple probability theory, however, shows the large risks of such a rule.

You might compute the probability of rejecting lots for other values of π besides .01 and .50. Obviously, there are many. But always remember these probabilities depend on which state of the world we assume to be true.

In the preceding illustration the sample size was small. In contrast, suppose our sampling process involved the selection of a random sample of 200 items from a lot of 5,000 units. The decision rule might then be: If 5 or more items are defective, accept the lot; otherwise reject the lot. We could calculate the risks of this decision rule as we did before, but the arithmetic would be arduous. It would involve the multiplication of 200 fractions. Fortunately, techniques have been developed to simplify such computations. This will be part of the subject matter of Chapter 6.

Now, however, suppose that we turn to other important concepts of probability theory; concepts which will be useful to us as theoretical tools in our study of statistical decision making.

5.5 Independent Repeated Trials

When repeated trials are independent rather than dependent, a special case of the conditional probability rule may be applied. This

special case is frequently encountered in statistics, but first may be illustrated by reference to a game of chance. For example, what is the probability of obtaining two heads in two tosses of an unbiased coin? Applying our conditional probability rule

$$P(H \text{ AND } H) = P(H) \times P(H|H).$$

That is, the probability of a sequence of two heads is equal to the probability of the first head multiplied by the probability of a second head given that the first toss is a head. Thus,

$$\begin{aligned} P(H \text{ AND } H) &= \frac{1}{2} \times \frac{1}{2} \\ &= \frac{1}{4}. \end{aligned}$$

You will observe that since the coin is unbiased, the probability of a second head is not influenced by the occurrence of the first head. The probability is $\frac{1}{2}$ in both cases. In other words, the conditional probability $P(H|H)$ is the same as the unconditional probability, $P(H)$. This always is the case when the outcomes under consideration result from independent trials.

We now can restate the conditional probability rule for the special case. When trials are independent,

$$P(A \text{ AND } B) = P(A) \times P(B).$$

The rule can be extended to more than two trials as follows:

$$P(A \text{ AND } B \text{ AND } C \text{ AND } \dots) = P(A) \times P(B) \times P(C) \times \dots$$

We are especially interested in the rule for independent trials when each trial can result in one of two mutually exclusive outcomes. For example, a coin can be a head or a tail; a television cabinet can be defective or nondefective; a person can be male or female; a family's income can be less than \$10,000 a year, or \$10,000 or more a year. In our discussion of repeated independent trials we will concern ourselves only with such outcomes.

The rule for independent trials is applicable in answering questions such as the following: What is the probability that in a simple random sample of 100 entries from an accounts payable ledger there will be 2 incorrect entries and 98 correct entries? Suppose that we are concerned with the system of recording ledger entries. Hence, we are dealing with an analytical study. Our population is, therefore, infinite—consisting of the ledger entries past, present, and future made under similar conditions. The probability of an entry being selected as the second unit in a sample is not conditional upon the selection of the first entry. When 1 entry is chosen, there still remain an unlimited number of entries.

Suppose that we reconsider an earlier illustration to indicate the use of the rule for independent trials in the evaluation of the risks of relying on a decision rule.

Example 5.13 A firm which has a policy of attempting to fill orders within a week's time has reasons to suspect that its shipping process has slowed down. In the past it has been determined that only 5 per cent of its orders were delayed more than a week—and this is regarded as satisfactory. Now, however, several customer complaints hint at the possibility that π , the proportion of delayed orders, has increased from its normal and satisfactory level of .05.

Note that this situation calls for an analytical study. It deals with a (shipping) process and, therefore, presents an infinite population.

Suppose that the firm plans to select a sample of 3 orders as a basis for deciding on whether or not to attempt to improve the shipping process. How will they use the information obtained from this sample? If none of the 3 shipments is delayed, the shipping process will not be checked; if 1 or more are delayed, the process will be investigated with an eye towards revising it. This is their statistical decision rule—but what is the risk of applying it? Actually there are many risks.

One is the risk of investigating the process because the sample shows 1 or more delayed orders, even though π has remained at its normal and satisfactory level of .05. If we assume that the 3 orders selected for the sample are filled independently, this risk may be computed by the use of the rule for independent trials. Thus, if $\pi = .05$, for a single order selected at random,

$$\begin{aligned} P(\text{DELAYED ORDER}) &= .05 \\ P(\text{UNDELAYED ORDER}) &= .95. \end{aligned}$$

Then the chance of not investigating the process is the chance of selecting 3 successive undelayed orders, namely,

$$\begin{aligned} &P(\text{UNDELAYED AND UNDELAYED AND UNDELAYED}) \\ &= P(\text{UNDELAYED}) \times P(\text{UNDELAYED}) \times P(\text{UNDELAYED}) \\ &= .95 \times .95 \times .95 \\ &= .857. \end{aligned}$$

It follows that the probability of investigating if $\pi = .05$ is $1 - .857 = .143$. This is the risk of checking a normal and satisfactory process.

Another risk is that the sample will lead to not investigating the shipping process when π actually has increased significantly. For example, suppose $\pi = .25$. Then,

$$\begin{aligned} P(\text{DELAYED ORDER}) &= .25 \\ P(\text{UNDELAYED ORDER}) &= .75 \end{aligned}$$

and the chance of not investigating the process is

$$.75 \times .75 \times .75 = .422,$$

while the chance of investigating is $1 - .422 = .578$.

Similar calculations indicate some of the other probabilities of investigating and of not investigating the process are as follows:

π	$P(\text{Investigating})$	$P(\text{Not Investigating})$
.00	.000	1.000
.05	.143	.857
.10	.271	.729
.15	.386	.614
.20	.488	.512
.25	.578	.422

Check one or two of these values. Also, notice that these probabilities are all characteristics of the decision rule under consideration. To evaluate that rule you should consider these chances or risks. Do they seem economical or not? Why or why not?

The last illustration affords an opportunity for us to make note of an important thought. You can see that the probability of investigating the process increases as π increases. Generally speaking, this means that the more the shipping process has deteriorated, the better our chances of taking the right action—of checking the process. Thus, for $\pi = .10$, this probability is .271, while for $\pi = .25$, it is .578. For less and less desirable states of the world the probability of taking the right action gets higher and higher, and the risk of making an incorrect decision gets lower and lower. This is true despite the fact that the sample size is quite small. What would a larger sample accomplish? It would mean smaller risks for the different values of π .

5.6 The Binomial Formula

In the preceeding section we used the rule for independent trials to evaluate the risks in relying on sample information from an infinite population. While the method of calculation we used to compute these probabilities is relatively simple for small samples, a more generalized procedure would facilitate the determination of probabilities of outcomes for any size of sample.

Such a generalized procedure is afforded by mathematical probability. It takes the form of a formula called the *binomial formula*. This formula simply represents an extension of the rule for independent trials. It also is related to the binomial theorem or binomial expansion, a relatively simple mathematical relationship of importance in algebra. Perhaps you remember studying the binomial; if so, you will recognize the relationship between it and what we are studying here. If not, don't worry about it—just remember that the formula we are studying is called the binomial.

Let us indicate the derivation of this formula. Suppose that we have n independent trials, such as n selections from an infinite population, and that each of the n trials can result in only one of two outcomes. To keep a concrete example in mind, envision a production process producing n parts each of which may be either defective (D) or nondefective (N) and assume these parts are produced independently.

What is the probability of a particular sequence? If we let

$$\begin{aligned} P(\text{ONE DEFECTIVE PART}) &= \pi \\ P(\text{ONE NONDEFECTIVE PART}) &= 1 - \pi, \end{aligned}$$

then by the rule for independent trials, the probability of the sequence D, N, D is

$$\pi \times (1 - \pi) \times \pi,$$

or, in short,

$$\pi^2(1 - \pi).$$

Similarly, the probability of the sequence N, D, D is

$$(1 - \pi) \times \pi \times \pi = \pi^2(1 - \pi),$$

while that of the sequence D, D, N is

$$\pi \times \pi \times (1 - \pi) = \pi^2(1 - \pi).$$

You should note that all three of these sequences refer to 1 nondefective and 2 defectives—although they refer to their occurrence in different *orders*. It is no surprise, then, that the probability of each is $\pi^2(1 - \pi)$. Now, it is important to recognize that the outcome D occurs twice in each sequence, and, hence, the exponent on π , the probability of D occurring, is 2. Since N occurs once the exponent on $(1 - \pi)$, the probability of N occurring, is 1.

Suppose, now, that we are not interested in the order of occurrence and that we simply need to know the probability of 1 nondefective and 2 defectives in three trials. This is the probability of observing $N D D$ or $D N D$ or $D D N$ and equals

$$\pi^2(1 - \pi) + \pi^2(1 - \pi) + \pi^2(1 - \pi) = 3\pi^2(1 - \pi)$$

because they are mutually exclusive possibilities.

Each of the terms, 3 , π^2 , and $(1 - \pi)$ has an interpretation. We have already noted that the exponents on π and $(1 - \pi)$ represent the frequency of D and N in a sequence. A coefficient, such as the 3 , indicates the number of sequences in which, say, 1 nondefective and 2 defectives can occur.

Thus, if we were interested in the probability of 2 defectives and 2 nondefectives, we know that we must have $\pi^2(1 - \pi)^2$ times some coefficient. What does this coefficient equal? There are six orders of two D 's and two N 's, namely,

$$\begin{array}{cc} D D N N & N D D N \\ D N D N & N D N D \\ D N N D & N N D D \end{array}$$

and hence the coefficient is 6. Thus, we may write

$$P(2 \text{ DEFECTIVES AND } 2 \text{ NONDEFECTIVES}) = 6\pi^2(1 - \pi)^2.$$

The binomial is a generalized formula for computing the probability of a given series of outcomes *in any order*. Thus, consider n trials in which

$$\begin{array}{l} r \text{ are defective, and} \\ n - r \text{ are nondefective.} \end{array}$$

The probability of any such sequence is

$$\pi^r(1 - \pi)^{n-r}$$

since the exponent on π refers to the number of outcomes that are defective and the exponent on $(1 - \pi)$ refers to the number of outcomes that are nondefective.

Furthermore, the probability of r defectives and $n - r$ nondefectives *in any order* is $\pi^r(1 - \pi)^{n-r}$ multiplied by the number of sequences in which r defective and $n - r$ nondefective can occur. This number of sequences may be expressed in terms of factorials as

$$\frac{n!}{r!(n - r)!}.$$

Factorials, as you should recall, were defined in Chapter 4 where a similar formula was developed for determining the number of possible samples that could be drawn from a population. With $n = 4$ and $r = 2$, we find

$$\frac{4!}{2!2!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1 \times 2 \times 1} = 6,$$

a result which we also found earlier by direct enumeration.

Thus, if we write $P\left(\frac{r}{n}\right)$ for the probability of r outcomes of a specified type (such as defectives) in n trials,

$$P\left(\frac{r}{n}\right) = \frac{n!}{r!(n - r)!} \pi^r (1 - \pi)^{n-r}.$$

This is the binomial formula. Remember that

- n = number of trials,
- r = number of specified outcomes,
- $n - r$ = number of other outcomes,
- π = probability of a specified outcome in a single trial.

The binomial, as noted, is a general formula for computing probabilities for independent trials, and as such is useful in statistics for computing risks of relying on sample information when the sample observations are independently derived attributes.

Example 5.14 Each hour a production manager selects a sample of 5 parts produced by a machine. If more than one of the parts is defective, he stops the process and resets the machine; if none or one are defective, he permits it to continue as is. What is the risk of stopping the process when π is only .02 (a quality level the firm considers satisfactory)? The binomial provides the answer. The probability of not stopping the process is the probability of finding 0 or 1 defectives in 5 trials. Hence, we should let

$$n = 5, r = 0, \pi = .02,$$

thus finding, for the probability of 0 defectives,

$$\begin{aligned} P\left(\frac{0}{5}\right) &= \frac{5!}{0! 5!} (.02)^0 (.98)^5 \\ &= (.98)^5 \\ &= .904. \end{aligned}$$

(Remember that $0! = 1$ and $(.02)^0 = 1$.)

Similarly, for the probability of 1 defective,

$$\begin{aligned} P\left(\frac{1}{5}\right) &= \frac{5!}{1! 4!} (.02)^1 (.98)^4 \\ &= 5(.02)(.98)^4 \\ &= .092. \end{aligned}$$

These two results may be combined to show the probability of 0 or 1 defective (which is the probability of not stopping the process):

$$.904 + .092 = .996.$$

Thus, the risk of stopping the process if $\pi = .02$ is only $1 - .996 = .004$. Does this mean that the production manager's rule is a good one? Why or why not?

Unfortunately, even the binomial formula involves more and more complicated computations for larger and larger sample sizes. Therefore, in the next chapter we will return to the binomial and see how these calculations of probabilities may be approximated and, thus, simplified further.

5.7 Expected Values

In this chapter we have considered ways of measuring and computing probabilities of various outcomes in a random experiment or a series of random experiments. In this particular section we shall deal with an allied topic—*expected value*. This is the average value that would occur if the random experiment were to be repeated indefinitely.

For example, suppose a die is thrown and the face value observed. The average value observed after many throws is what is meant by the expected value. What is this expected value? Assuming each side equiprobable, we know that each face value—1, 2, 3, 4, 5, and 6—is equally likely; that is, the probability of each is $\frac{1}{6}$. By assumption, then, in the long run each of these values would occur $\frac{1}{6}$ of the time. The average value is thus,

$$\frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = \frac{21}{6} = 3.5,$$

and the 3.5 is called the expected value of the random variable under consideration.

Generally speaking, the expected value of any random variable may be computed by (a) multiplying each possible value the variable may assume by its probability of occurrence, and (b) summing these products. This is the method used in the preceding paragraph—each value (1, 2, 3, 4, 5, and 6) was multiplied by its probability of occurrence ($\frac{1}{6}$), and these results were added together.

Example 5.15 Suppose that on the basis of his past experience, a newsdealer estimates the probabilities of selling 7, 8, 9, or 10 magazines per hour as follows:

$$\begin{aligned}P(7) &= .2 \\P(8) &= .4 \\P(9) &= .3 \\P(10) &= .1.\end{aligned}$$

The expected number of magazines is then

$$7(.2) + 8(.4) + 9(.3) + 10(.1) = 8.3.$$

This is an average value—an average that might be expected in the long run.

Example 5.16 A manufacturer receives shipments of a special part in lots of 50. He uses the following decision rule to decide whether or not to inspect the entire lot:

Select a simple random sample of 5 parts. If none of the 5 selected parts is defective, do *not* inspect the remaining 45 parts. If 1 or more are defective, inspect the remaining 45 parts.

The manufacturer desires to know the expected amount of inspection in order to estimate the cost of inspection under this decision rule.

The total amount of inspection for any one lot depends upon the action taken. Note the tabulation that follows:

Action	Number of Items Inspected		Total Number of Parts Inspected
	Sample	Remainder of Lot	
No further inspection...	5	0	5
Full inspection.....	5	45	50

For any lot, however, the total number of parts inspected is a random variable since it results from a random experiment of testing a sample of 5 parts. To find the expected amount of inspection it is, therefore, necessary to determine the probability of each action under a given assumption about π .

Let us assume that a shipment contains 5 defectives—i.e., π (the proportion of defectives) is $5/50$, or $.10$. Then under the decision rule

$$\begin{aligned}
 P(\text{NO FURTHER INSPECTION}) &= P(0 \text{ DEFECTIVES IN A SAMPLE OF } 5) \\
 &= \frac{45}{50} \times \frac{44}{49} \times \frac{43}{48} \times \frac{42}{47} \times \frac{41}{46} \\
 &= .577.
 \end{aligned}$$

It follows, therefore, that for any lot with $\pi = .10$, the probability of fully inspecting a lot is: $1 - .577 = .423$.

We can compute the expected amount of inspection:

$$\begin{aligned}
 E(\text{AMOUNT OF INSPECTION}) &= (5 \times .577) + (50 \times .423) \\
 &= 24.0.
 \end{aligned}$$

This means that if many lots with $\pi = .10$ are received, on the average 24.0 parts will be inspected.

However, you must remember that the expected amount of inspection depends on the probabilities of no inspection and of full inspection—and, hence, on the proportion of defectives in a lot. We have found the answer for $\pi = .10$, and similar calculations would enable us to determine the expected amount of inspection for other values of π . (For example, suppose $\pi = .00$. Then the probability of fully inspecting a given lot is 0. Why? Therefore, the expected amount of inspection is 5. Why?)

An expected value is a special type of arithmetic mean. That is, it is an average value. It is special only in the sense that it is the

average value of a random variable—the average of numbers which are the result of a chance experiment.

Thus, it must be stressed that the term “expected value” does not refer to what we expect to occur on any one trial—in any single random experiment. An expected value is an average of what would happen only if such an experiment were repeated over and over again (an unlimited number of times). Thus, to refer back to the fact that the expected value of the outcome of throwing a die is 3.5, we can see it is impossible to get a 3.5 on a single toss. Certainly this is not what we should “expect” on any one throw. The 3.5 simply indicates the average value that we would expect to occur in the long run.

It is customary in mathematics and statistics to use the symbol E to represent the term “expected.” Thus, if we let x refer to the face value that occurs when a die is thrown, $E(x)$ refers to the “expected value” of this experiment and $E(x) = 3.5$.

The concept of expected value is useful in many connections in probability theory and in its fields of application. In statistics, it is used primarily in connection with *evaluating sampling procedures* and, hence, in determining sound statistical decision procedures.

As has been stressed in this chapter, statistics such as the sample mean ($\bar{x}$), the sample variance (s^2), and the sample proportion (p) are random variables. Since they are random variables, we may apply the idea of expected value to them. This point will be discussed more fully in the next chapter.

Note for the present, however, that by introducing the concept of expected values we have introduced a *third* application of the average which is useful in statistics. Thus, μ represents one use for the average—it represents the average value in a population. A second use for the average is to summarize information in a sample—in this connection an average is denoted by $\bar{x}$. Thirdly, any expected value is an average—the long-run average value of a random variable. These uses for the concept of the average are all important, and they are all qualitatively different. Don't confuse them.

5.8 You Should Now Know That

A statistic is a property of a sample. It summarizes some information contained in the sample.

Statistics are variables. They vary from sample to sample. When they are properties of probability samples, they are random variables.

The way in which they vary can be predicted on the basis of probability theory.

Statistics of present interest and their symbolic names are:

- p : the sample proportion
- $\bar{x}$: the sample arithmetic mean
- s : the sample standard deviation
- s^2 : the sample variance

The theory of probability is a foundation stone of statistical theory and methods.

A numerical measure of probability is called a probability measure.

This measure may be interpreted as the proportion of times an outcome would occur in an infinite series of repeated trials.

Among the bases for estimating probabilities are (a) past experience, and (b) the assumption that all outcomes are equiprobable.

The basic properties of a probability measure are:

- a) Any probability can range only from 0 to 1,
- b) The sum of the probabilities of all possible outcomes equals 1,
- c) The probability of A or B equals the probability of A plus the probability of B , if A and B are mutually exclusive outcomes.

Repeated trials are dependent if they influence one another; they are independent if they do not influence one another in any way.

The probability that two outcomes occur in a series of repeated trials can be computed as follows:

- a) For two dependent trials this rule may be expressed as

$$P(A \text{ AND } B) = P(A) \times P(B|A),$$

where $P(B|A)$ is a conditional probability—the probability B occurs given that A has occurred.

- b) For two independent trials this rule becomes

$$P(A \text{ AND } B) = P(A) \times P(B)$$

because $P(B|A) = P(B)$ if A and B are outcomes in independent trials.

An expected value is an average value. It represents the average value that we would expect to occur in the long run. It does not represent a value that we expect to occur on a single trial.

5.9 You Should Now Be Able to Solve

1. The mean diameter of a shipment of 500 ball bearings is .248 inches. Under what circumstances would this average be a decision-

parameter (μ)? Under what circumstances would it be a statistic ($\bar{x}$)?

2. Mortality tables tell us, for example, the probability of a man alive at age 21 dying before his 22nd birthday. Why should such tables be revised periodically? Consider the definition of probability.

3. A statistics student has investigated the occurrence of fires in his street and has found out that, on the average, a fire occurs once in 5 years. A fire has occurred this year. He, therefore, advises his father not to purchase a fire insurance policy for the next 4 years. Should the student pass statistics? Why or why not?

4. Explain why each of the following is either true or false:

- a) If the probability of an employee being absent is .01 and the probability of an employee being late is .02, then the probability of an employee being either late or absent is .03.
- b) If the probability that a television cabinet will have a scratch is .001 and the probability that it will have a dent is .002, then the probability that the cabinet will have either a scratch or a dent is .003.
- c) The probability of a clerk misfiling a paper is .03. This means that we will find 3 errors if we examine a probability sample of 100 filed papers.
- d) If the probability of producing a defective product is .05, the probability that 2 successive items will both be defective is .10.
- e) A company manufactures 12-inch rulers. Some rulers, however, are slightly longer and some are slightly shorter. If the probability of a ruler being longer than 12 inches is .40, the probability of a ruler being shorter than 12 inches is .60.

5. On the basis of past experience, it is estimated that the probability of a machine breaking down during a week's time is .05. What is the probability that it will break down within 4 weeks? What is the probability that it will run for 2 weeks without breaking down? What assumptions did you make in answering these questions?

6. A common decision procedure to decide whether to accept or reject lots is the following: Inspect a sample of 5 articles selected at random from a lot of 50; if no article is defective, accept the lot; if one or more articles are defective, reject the lot.

What is the probability of accepting lots that contain 2 per cent defective? Six per cent defective? Ten per cent defective? On the basis of these calculations do you think that the decision rule offers protection against accepting lots of inferior quality?

7. Suppose the error rate in a posting operation is 3 per cent. If a sample of 10 postings is observed, what is the probability of finding no errors? Two errors? Five errors?

8. A sample of dwelling units is to be drawn by selecting a block at random and including every dwelling unit in the block in the sample. The blocks and the number of dwelling units in each are as follows:

Block	Number of Dwelling Units
A.....	10
B.....	9
C.....	12
D.....	14
E.....	8

What is the expected value of the sample size? How do you interpret this value?

5.10 You Will Also Find That

Some of the basic principles of the mathematical theory of probability, remarks on its applicability, and notes on its historical development are discussed in the following references:

CRAMÉR, HARALD. *The Elements of Probability Theory*, pp. 11–28. New York: John Wiley & Sons, Inc., 1955.

POLYA, G. *Mathematics and Plausible Reasoning*, Vol. II: *Patterns of Plausible Inference*. Princeton, N.J.: Princeton University Press, 1954.

Probability and its relationship to statistics are treated by:

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics: A New Approach*, pp. 309–41. Glencoe, Ill.: The Free Press, 1956.

HIRSCH, WERNER Z. *Introduction to Modern Statistics*, pp. 66–90. New York: Macmillan Co., 1957.

DUNCAN, ACHESON J. *Quality Control and Industrial Statistics*, pp. 13–26. Rev. ed. Homewood, Ill.: Richard D. Irwin, Inc., 1959.

Sampling Distributions

6.1 Sampling Distributions Defined

We now have discovered that statistics (measures computed from sample data) can be used in place of decision-parameters, when the latter are unknown. We have also seen that risks are connected with this procedure since the value of a statistic will vary depending on the particular sample selected. Whenever the sample which a statistic represents is a probability sample, this variation is random and is to some extent predictable. It is, so to speak, governed by the laws of chance and the theory of probability. Therefore, in the preceding chapter, we acquainted ourselves with some basic probability theory and used this theory to evaluate risks of relying on sampling procedures. Our discussions, however, were limited to small-size samples and to situations where the observations made on the characteristic of interest were attributes (as opposed to variates). We shall now investigate the *pattern of sampling variability* as an extension of our previous discussions.

First, let us consider what statisticians mean by the expression “pattern of sampling variability.” It refers to the distribution of all possible values that some given statistic can assume, plus the respective probabilities of those values occurring. Such a distribution is called a *sampling distribution* or, synonymously, a *probability distribution*.

We will use sampling distributions to summarize the probabilities of different values of a statistic occurring. Their use will enable us to calculate risks of relying on such statistics. As you will see, it is somewhat easier to compute risks for large samples by the use of sampling distributions than by the simple methods discussed in the last chapter.

Now let's look at the illustration that follows to discover some general characteristics of sampling distributions.

Example 6.1 Consider a hypothetical small town in which all the cars have license plates numbered serially. That is, one car has the plate number 1, another has 2, still another has 3, etc.

Suppose that there were 6 cars in town. A statistician is to select a simple random sample of 2 cars. He defines a statistic, g , as the highest serial number in his sample. What values of g are possible, and what are their probabilities of occurring? To answer this question suppose we list all possible samples, their probabilities of occurrence, and the value of g associated with each sample.

Sample	Probability	g
1,2.....	$\frac{1}{15}$	2
1,3.....	$\frac{1}{15}$	3
1,4.....	$\frac{1}{15}$	4
1,5.....	$\frac{1}{15}$	5
1,6.....	$\frac{1}{15}$	6
2,3.....	$\frac{1}{15}$	3
2,4.....	$\frac{1}{15}$	4
2,5.....	$\frac{1}{15}$	5
2,6.....	$\frac{1}{15}$	6
3,4.....	$\frac{1}{15}$	4
3,5.....	$\frac{1}{15}$	5
3,6.....	$\frac{1}{15}$	6
4,5.....	$\frac{1}{15}$	5
4,6.....	$\frac{1}{15}$	6
5,6.....	$\frac{1}{15}$	6

We can rearrange this tabulation by summarizing the values of g and their probabilities of occurrence.

g	Probability
2.....	$\frac{1}{15}$
3.....	$\frac{2}{15}$
4.....	$\frac{3}{15}$
5.....	$\frac{4}{15}$
6.....	$\frac{5}{15}$

This table represents a probability distribution or sampling distribution of the statistic g if there are 6 cars in town. It tells us, for example, that if there were 6 cars in town and we were to take a large number of samples, 4 out of 15 times the largest number in our sample would be a 5, 1 out of 15 times the largest number would be a 2, etc.

Notice that in this example the sampling distribution referred to a specific statistic (g), to a specific method of sampling (simple random), to a specific sample size ($n = 2$), and to specific conditions in the population (6 cars in town). If we had varied any of these conditions, say, by considering a sample of 5 or by assuming the popu-

lation contained 100 cars, we should get a different sampling distribution. Therefore, you should remember that whenever we speak of a sampling distribution we must speak of it with reference to:

1. A given statistic,
2. A method of sampling,
3. A sample size, and
4. A specific population.

The method of determining a sampling distribution that we have used in the last example is completely valid. However, it is not very effective from a practical point of view. For example, if the population is infinite, it is, of course, impossible to write down all possible samples of a given size—there are an infinite number. Even if the population contains a finite number of observations, enumeration usually is impractical. Thus, there are 75,287,520 possible samples of size 5 that might be drawn from a population of 100. You could spend your entire college career just computing all possible values of a given statistic. Therefore, in this chapter we must develop more sophisticated methods of determining the sampling distribution of a statistic.

6.2 The Sampling Distribution of p

The sample proportion (p) is one of the statistics used frequently in statistical decision problems. Let's consider its sampling distribution. This distribution presents all possible values of the sample proportion and the probabilities of their occurrence.

In the preceding chapter we developed an analytical tool—the binomial formula—to calculate the probability of r occurrences of a given attribute in a sample of n items. However, p , the sample proportion, is by definition r/n . For example, if in a sample of 5 items, 2 are defective, $p = .40$. Hence, the probability of 2 defectives in a sample of 5 items is the same as the probability of $p = .40$. Therefore, we may use the binomial formula to determine the probability of all possible values of p . Remember that the sampling distribution of p is expressed exactly by the binomial if and only if we are concerned with simple random samples from an infinite population of attributes. Furthermore, the distributions we derive must always refer to a fixed sample size and specified conditions about the population (a fixed π).

Let's restate the binomial formula:

$$P\left(\frac{r}{n} = p\right) = \frac{n!}{r! (n-r)!} \pi^r (1-\pi)^{n-r}.$$

In applying it for a given n and π we let r equal $0, 1, \dots, n$ in order to get a set of $n + 1$ probabilities, each of which is associated with a certain r or p . This set of probabilities is called the sampling distribution of p . However, for different values of n or π we obtain different probability distributions.

Our main task in this section is, therefore, to consider the effect of different values of n and π upon the sampling distribution of p . To start, let's fix n at 5 and see what the probabilities of all possible values of p are when $\pi = .05, .25, .50, .75$, and $.95$. By successive application of our binomial formula for $r = 0, 1, 2, 3, 4$, and 5 , for $n = 5$, and for the appropriate values of π , we find the following five sampling distributions:

p	$\pi = .05$	$\pi = .25$	$\pi = .50$	$\pi = .75$	$\pi = .95$
.00.....	.774	.237	.031	.001	.000
.20.....	.204	.395	.156	.015	.000
.40.....	.021	.264	.313	.088	.001
.60.....	.001	.088	.313	.264	.021
.80.....	.000	.015	.156	.395	.204
1.00.....	.000	.001	.031	.237	.774

The five probability distributions of p , tabled above, are shown in Figure 6.1. Look at them for a minute or two so that we can discuss them with a clear picture in mind. First of all, notice that the sampling distribution of p for $\pi = .50$ is *symmetrical*. In short, the left side is a mirror image of the right, and vice versa. However, the other four distributions are skewed. The distributions for $\pi = .05$ and $\pi = .25$ are skewed to the right. Alternately, the distributions for $\pi = .75$ and $\pi = .95$ are skewed to the left. In addition, we notice that those distributions for $\pi = .05$ and $\pi = .95$ are more skewed than those for $\pi = .25$ and $\pi = .75$. These observations permit the following generalizations:

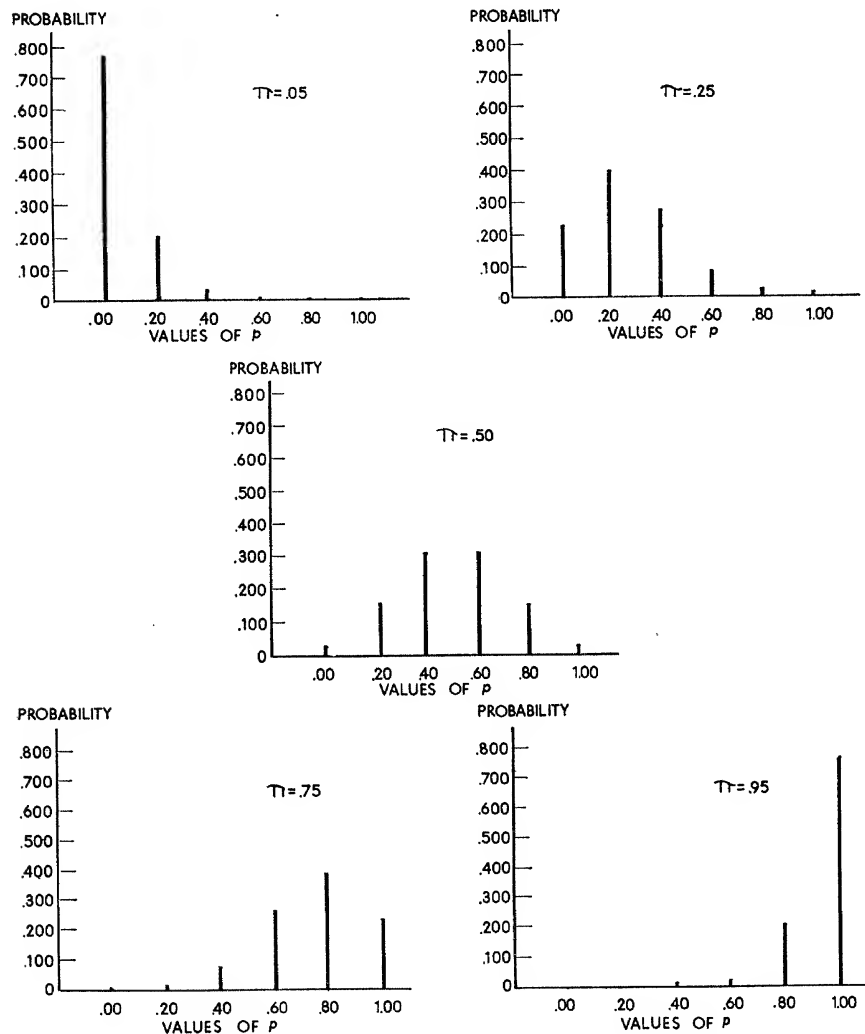
1. The sampling distribution of p for $\pi = .50$ is always symmetrical no matter what the sample size.
2. The sampling distribution of p is skewed to the right if π is less than .50. It is skewed to the left if π is more than .50.
3. The further π departs from .50, the greater the skewness of the sampling distribution of p , for a given sample size.

In addition, notice that the distributions of $\pi = .25$ and $\pi = .75$ are mirror images of each other. Similarly, the distributions for $\pi = .05$ and $\pi = .95$ are mirror images. Why?

It is important to observe that the most probable values of p

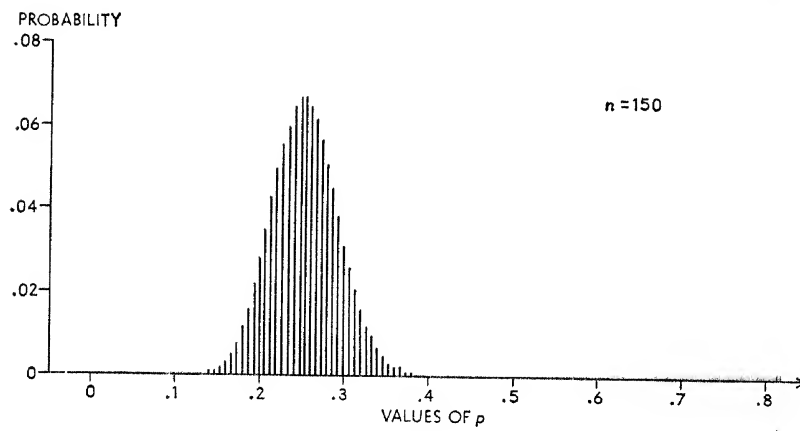
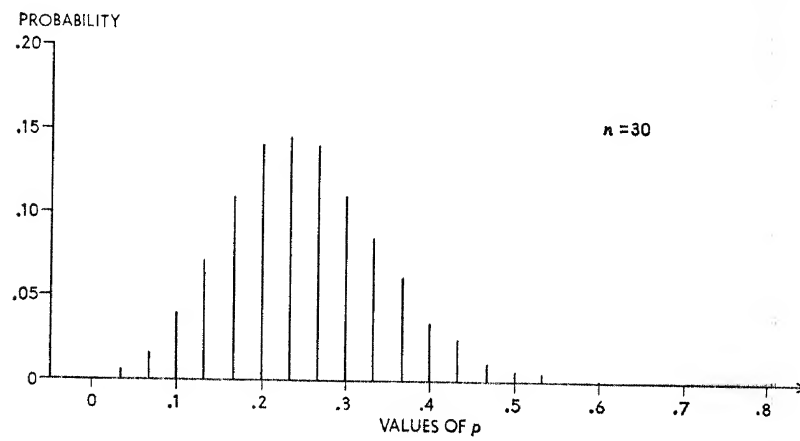
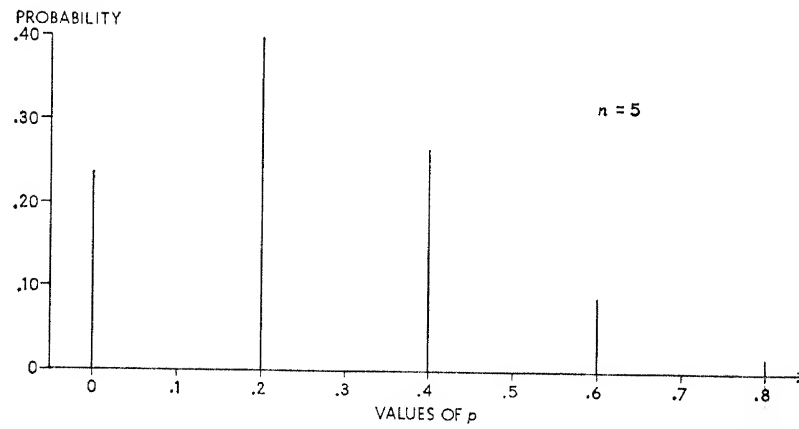
always are in the neighborhood of the value of π . Thus, if $\pi = .05$, low values of p are the most likely and high values of p are the least likely. On the other hand, if $\pi = .95$, high values of p are the more likely—low values of p may occur, but their chances are small.

Figure 6.1



Now let's consider what happens to the sampling distribution of p as we increase the sample size. In Figure 6.2 sampling distributions of p are portrayed for sample sizes 5, 30, and 150 when $\pi = .25$. The probabilities needed to construct these charts were derived by using the binomial formula repeatedly. One thing you notice im-

Figure 6.2



mediately is that there are more vertical lines on the graphs as n increases— p can attain a wider variety of values when n is large. More important, the concentration about the value of π gets more and more pronounced as the sample size increases. That is, the probability that p will be approximately equal to π gets greater and greater.

Lastly, we should observe that as the sample size increases, the sampling distribution of p becomes more nearly symmetrical and more nearly bell-shaped. In fact, as n gets larger, the distribution of p tends to resemble that of a *normal curve* more and more. This observation can be proved mathematically. That is, for sufficiently large simple random samples, it may be shown that the sampling distribution of p for any given value of π may be well approximated by a normal curve.

If the sample size is large enough, therefore, we can calculate probabilities of values of p by means of the normal curve rather than by the binomial formula. But, what do we mean by “large enough”? We know that for small samples, the binomial distribution is symmetrical when π is .50, but it becomes more and more skewed as π departs from .50 in either direction. This indicates that the more π departs from .50, the greater the sample size needed for the normal curve to offer a satisfactory approximation of the binomial distribution. The table that follows presents general rules suggested by one statistician to decide whether n is “sufficiently large.” It indicates the smallest value of n for which the normal approximation should be used if π is a certain value.

<i>If π Equals</i>	<i>Use the Normal Approximation Only if n Is at Least Equal to:</i>
.5	30
.4 or .6	50
.3 or .7	80
.2 or .8	200
.1 or .9	600
.05 or .95	1,400

Reprinted with permission from William G. Cochran, *Sampling Techniques* (New York: John Wiley & Son, Inc., 1953), p. 41.

6.3 Characteristics of the Sampling Distribution of p

In the last section we developed a second important use of the normal curve—as an approximation of the sampling distribution of p . Heretofore, in Chapters 2 and 3, we considered the normal curve as a population model. At that time, we saw that areas under the normal curve—in other words, the relative frequency with which

observations were contained in a given interval—could be determined if we knew the population mean and standard deviation. Now, we similarly could determine the relative frequency with which sample proportions fall within a given interval if we knew the mean and standard deviation of the sampling distribution of p . For example, we would be able to say that approximately 68 per cent of all possible sample proportions would be included in an interval of plus and minus one standard deviation about the mean of the sampling distribution. The 68 per cent, in this case, represents the relative frequency of occurrence of outcomes of a random variable—the sample proportion. Therefore, in keeping with the relative frequency definition of probability, the .68 may be also interpreted as a probability. Thus, areas, relative frequencies, and probabilities are equivalent concepts when the normal curve is used to describe a sampling distribution.

But to measure these probabilities we must first know the mean and standard deviation of the sampling distribution of p . Let us first consider the mean. We already have mentioned this type of arithmetic mean in another connection. In Chapter 5 we pointed out that the expected value of any random variable was the average or the arithmetic mean of the values that would occur in the long run. At present we are interested in a particular random variable—the sample proportion, p . Therefore, we are interested in the expected value of p —the arithmetic mean of the sampling distribution of p . Symbolically this is represented by $E(p)$.

Example 6.2 Consider the sampling distribution of p for $n = 4$ and $\pi = .30$. The results were derived through the use of the binomial formula.

p	Probability
.00	.2401
.25	.4116
.50	.2646
.75	.0756
1.00	.0081

To calculate the expected value of p we must multiply each possible value of p by its probability of occurrence, and then add the products.

$$\begin{aligned}
 E(p) &= (.00)(.2401) + (.25)(.4116) + (.50)(.2646) \\
 &\quad + (.75)(.0756) + (1.00)(.0081) \\
 &= .0000 + .1029 + .1323 + .0567 + .0081 \\
 &= .30
 \end{aligned}$$

This value, .30, represents the average of the values of p that would occur in the long run—it is the “expected value” of p if a simple ran-

dom sample of size 4 were selected from a population in which $\pi = .30$.

Now, note that in our example, $E(p) = .30$ and that $\pi = .30$. This is no coincidence—it is a general fact which can be proved mathematically. That is, whenever we use simple random sampling,

$$E(p) = \pi.$$

What does this imply? It tells us that in the long run the average sample proportion will be equal to the population proportion—no matter what the value of this population proportion.

Of course, this long-run business may not mean too much to us. After all, in practical applications, we ordinarily will select only one sample. Therefore, in addition to the expected value of p , we would like to find out something about the variability of p . Would it vary a great deal from sample to sample if we selected a series of samples? By how much, on the average, will any one value of p depart from the expected value of p ?

To answer these questions we may turn to the *standard deviation* of p or, more fully, the standard deviation of the sampling distribution of p . In short, we are going to consider a single measure of the variability of p . We will use the symbol σ_p for this measure. In addition to being called the standard deviation of p , this measure usually is called the *standard error of p* . We will use the terms interchangeably. Since the standard error of p measures the variability of all possible sample proportions about the population proportion (the mean of the sampling distribution of p), it is a measure of the reliability of sample results. Obviously, the smaller the standard error of a proportion, the closer the possible sample proportions will be to the population proportion and, hence, the more reliable will be the sample results.

For any given sampling distribution, σ_p or σ_p^2 (the variance of p) can be computed directly from the sampling distribution. The variance of p represents the average of the squared deviations of all possible sample proportions about the population proportion. However, since p is a random variable, the squared deviation of p about π also is a random variable. But, we have called the average value of a random variable, the expected value. Therefore, we can say that σ_p^2 is the expected value of the squared deviations of p about π or

$$\begin{aligned}\sigma_p^2 &= E(p - \pi)^2 \\ &= \Sigma[(p - \pi)^2 \times P(p)].\end{aligned}$$

Thus to obtain the variance of a random variable such as p , we must multiply each squared deviation by its probability of occurrence, and then add the products. Let's now apply this formula to the sampling distribution of p in Example 6.2 where n was 4 and π was .30:

p	$P(p)$	$(p - \pi)$	$(p - \pi)^2$	$\Sigma[(p - \pi)^2 \times P(p)]$
.00	.2401	-.30	.0900	.021609
.25	.4116	-.05	.0025	.001029
.50	.2646	+.20	.0400	.010584
.75	.0756	+.45	.2025	.015309
1.00	.0081	+.70	.4900	.003969
Total . . .				.052500

$$\sigma_p^2 = .0525$$

$$\sigma_p = .23 .$$

This is a tedious method of computing σ_p . However, the difficulty is eliminated by a simple formula which can be derived mathematically. It states

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}} .$$

Substituting the values of π and n from the preceding computation, we obtain the same result

$$\begin{aligned} \sigma_p &= \sqrt{\frac{(.30)(1 - .30)}{4}} \\ &= \sqrt{\frac{.21}{4}} \\ &= \sqrt{.0525} \\ &= .23 . \end{aligned}$$

This value of σ_p is a measure of the variability in a sampling procedure. It describes the variability among values of p (for samples of size n) that would result with repeated simple random sampling from an infinite population in which π equals .30.

Let us now consider some important aspects of σ_p . You will notice from the formula for σ_p that it depends on n , the sample size, and π , the population proportion. Notice that it varies inversely with n —that is, for larger and larger values of n , σ_p is smaller and smaller.

This expresses an obvious but very important fact: *If we increase the sample size, the distribution of sample proportions will become less and less variable.* In other words, the reliability of sample results increases as the sample size increases.

For example, consider the following values of σ_p for varying sample sizes when the population proportion, π , equals .5:

n	σ_p
25.....	.1000
100.....	.0500
500.....	.0224
1,000.....	.0158
10,000.....	.0050

Thus, as n gets larger, σ_p gets smaller.

Now notice that the standard error of p varies directly with the value of $\pi(1 - \pi)$. Below some values of $\pi(1 - \pi)$ are tabulated for various values of π :

π	$\pi(1 - \pi)$
1.00.....	.00
.90.....	.09
.70.....	.21
.50.....	.25
.30.....	.21
.10.....	.09
.00.....	.00

You see that $\pi(1 - \pi)$ is at a maximum when $\pi = .50$. As π departs from .50 in either direction, $\pi(1 - \pi)$ gets smaller.

The population proportion, π , is related to the amount of variability in such a population. For example, if all elementary units have, or do not have, a given attribute, π equals one or zero, respectively. In either case all observations are the same and there is no variability. However, the more variable a population, the closer the population proportion will be to .50. The most variable population of attributes is one for which $\pi = .50$. Thus, the more variable the population, the larger will be the value of $\pi(1 - \pi)$.

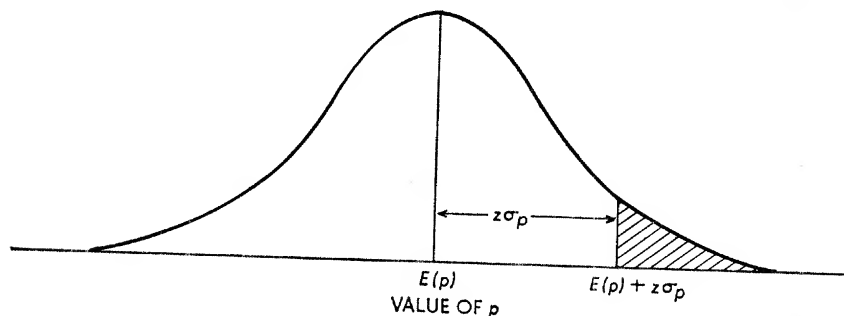
In summary, the larger the sample size or the less variable the population, the smaller σ_p will be. The smaller σ_p , the greater is the reliability in making decisions on the basis of sample proportions.

Now that we have determined the mean and the standard deviation of the sampling distribution of p , we can use this information to measure sampling risks when the normal curve approximates the binomial distribution as the sampling distribution of p .

6.4 The Normal Curve and Sampling Risks

We now know that for simple random samples from an infinite population, the normal curve may be used to describe the sampling distribution of p when the sample size is sufficiently large. In addition, we know how to determine the values of $E(p)$ and σ_p , the pertinent characteristics of the sampling distribution of p needed to determine areas under the normal curve. A table in Appendix B, to which we have referred before, gives areas under the normal curve. We shall use this table often to measure probabilities and, hence, should become proficient in its use.

The table in Appendix B tells us the area that is contained in one tail of the normal curve when we go out a given number of standard deviations from the mean. Thus, consider the normal curve as an approximation of the sampling distribution of p . With reference to the diagram below, the table of areas under the normal curve gives us the area in the tail (shaded portion) when we go z standard errors of p from $E(p)$, or π :

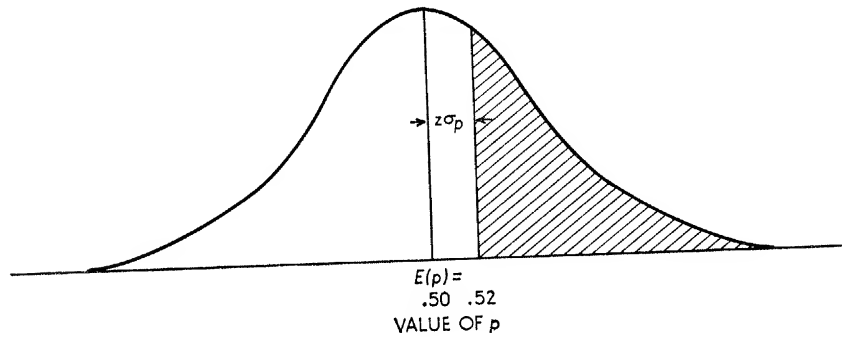


It is, therefore, necessary to specify z in order to use the table of areas. This quantity, z , is called the *normal deviate*. It represents the distance from the mean of the normal curve to a particular value of p , in multiples of the standard deviation. Symbolically z may be represented as

$$z = \frac{p - \pi}{\sigma_p}.$$

The distance between p and the mean of the distribution is $(p - \pi)$. It is divided by σ_p to determine how many standard deviation units it represents.

Example 6.3 What is the probability that a sample proportion will exceed .52 if it is computed from a simple random sample of size 100 drawn from an infinite population for which $\pi = .50$? The area in question is represented by the shaded portion in the diagram below:



To use the table of areas, we must determine z :

$$z = \frac{p - \pi}{\sigma_p}.$$

We first compute σ_p :

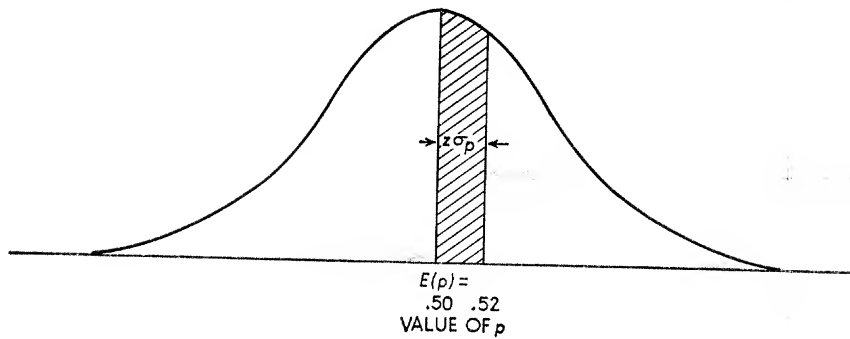
$$\begin{aligned}\sigma_p &= \sqrt{\frac{\pi(1 - \pi)}{n}} \\ &= \sqrt{\frac{.50 \times .50}{100}} \\ &= .05.\end{aligned}$$

Then we substitute in the expression for z :

$$\begin{aligned}z &= \frac{.52 - .50}{.05} \\ &= \frac{.02}{.05} = .40.\end{aligned}$$

The table of areas states that the area in question is .3446. Therefore, the probability of getting a sample proportion equal to .52 or more is .3446, if $\pi = .50$ and $n = 100$.

At times, we shall be interested in the area in the central portion of the curve rather than the area in the tail. For example, what would be the probability that a sample proportion computed from a simple random sample of 100 would fall between .50 and .52 if $\pi = .50$? The area in question is shaded in the diagram that follows:



We have found the area in the tail to be .3446. But, the area in half the normal curve, because of its symmetry, is .5000. Hence, the area in the central portion is

$$.5000 - .3446 = .1554,$$

and this represents the probability that p would fall between .50 and .52 if $\pi = .50$.

You might also verify that the probability that p would be less than .52 is

$$.5000 + .1554 = .6554,$$

or

$$1 - .3446 = .6554.$$

Why?

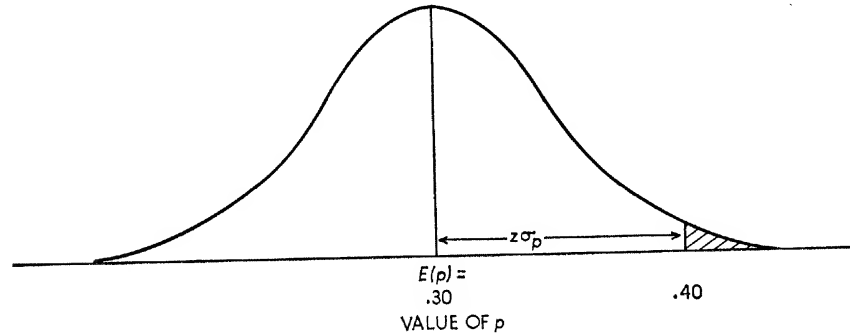
Our main interest in probability is as a measure of risk in relying on samples. Let's look at some illustrations of how the normal curve can be of use in this connection.

Example 6.4 A corporation has been contemplating the installation of an electronic computer to save time on paper work. The controller opposes the installation of the new equipment on the ground that paper work accounts for only 30 per cent of employees' work time. His assertion is challenged, and a study is undertaken to decide whether or not to install the equipment.

A statistical decision procedure is formulated. A simple random sample of 100 different time periods is to be selected. The worker to be observed at each time period is chosen at random—the same worker may be selected for observation more than once. At each time period a worker is reported to be doing or not to be doing paper work. The sample proportion p is the proportion of time periods workers are engaged in paper work. It is agreed that if p is found to be .40 or larger, the corporation will install electronic equipment; if p is less than .40, the equipment will not be installed.

What is the risk that the company will install the new equipment even though the controller's assertion that $\pi = .30$ is correct? The problem stated statistically is to determine the probability of obtaining a value of p equal to or greater than .40 if $\pi = .30$ and $n = 100$.

We can use the normal curve as an approximation of the sampling distribution of p since a sample size of 100 is sufficiently large for $\pi = .30$. The probability is measured by the shaded area in the diagram below:



To find z by the formula

$$z = \frac{p - \pi}{\sigma_p},$$

we note that

$$\begin{aligned}\sigma_p &= \sqrt{\frac{\pi(1 - \pi)}{n}} \\ &= \sqrt{\frac{(.30)(.70)}{100}} \\ &= .046\end{aligned}$$

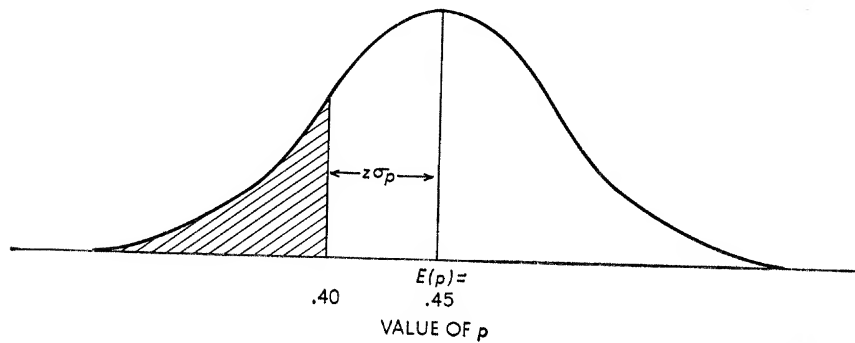
and, hence,

$$\begin{aligned}z &= \frac{.40 - .30}{.046} = \frac{.10}{.046} \\ &= 2.17.\end{aligned}$$

When $z = 2.17$, the area in the tail of the normal curve is .0150. In other words, the probability that p will equal or exceed .40 and, hence, that the company will install the equipment is .015 if $\pi = .30$. Thus, 15 times in a thousand, the corporation would be in error in using a decision rule such as the one suggested, if π was .30. Do you think this risk is too great?

Example 6.5 To illustrate further the use of the normal curve to measure risks, we continue with the same problem situation and sample size given in the preceding example. Suppose that π is .45. What is the probability that the corporation will not install electronic equipment under the same decision rule? Remember that the

equipment will not be installed if p is less than .40. Stated statistically, we are interested in determining the probability of obtaining sample proportions less than .40 from simple random samples of 100 if $\pi = .45$. The area in question is the shaded area in the diagram below:



To find z , we note that

$$z = \frac{p - \pi}{\sigma_p}$$

and that

$$\begin{aligned}\sigma_p &= \sqrt{\frac{\pi(1 - \pi)}{n}} \\ &= \sqrt{\frac{(.45)(.55)}{100}} \\ &= .050.\end{aligned}$$

Hence,

$$\begin{aligned}z &= \frac{.40 - .45}{.050} \\ &= \frac{-.05}{.050} \\ &= -1.00.\end{aligned}$$

Don't let the minus sign upset you. It merely indicates that p is less than π . Since the normal curve is symmetrical, we can ignore the sign and consult the table of areas for $z = 1.00$. We find this probability to be .1587.

Thus, in following its decision rule, about 16 times in a hundred the company would commit the error of not installing electronic equipment when actually its paper work amounted to 45 per cent of work time.

We have used the normal distribution to measure sampling risks in problems where the sample proportion was the criterion for decision. These risks could have been computed by use of the binomial

formula. However, the use of the normal curve greatly simplified the calculations necessary for the measurement of these risks.

6.5 Finite Twofold Populations

We have seen that the binomial gives us an *exact* distribution of the sampling distribution of p only when the proportions are computed from simple random samples selected from *infinite populations*. In addition, we know that as the size of such samples from infinite populations increases, the sampling distribution of p is approximated by the normal curve.

Although infinite populations of attributes occur frequently in statistical practice, so do finite populations. The characteristics of a lot of manufactured product, if each item's quality is measured as defective or nondefective, serves as one example. The population of homeowners versus renters in a town involves another. And in fact, any time we conduct an enumerative study we have another example of a finite population. In such instances the assumptions used to develop the binomial do not apply since the probability of one item being selected for a sample is not independent of, but conditional upon, the items that were selected before it.

The basic question is, then, how can we determine the sampling distribution of p when we select a simple random sample from a finite population? Is there a formula which will give us these probabilities just as the binomial serves us for infinite populations? The answer is "yes." A formula can be derived to determine the exact sampling distribution of p from a finite population.¹ Such a formula could be developed by the application of the conditional probability rule in the same way that the binomial formula was derived by the application of the rule for repeated independent trials. However, to develop this formula here would carry us into the small print of the study of statistics, and beyond the range of the present treatment.

Some characteristics of the distribution of p from a finite population, however, are of interest to us. In the first place, when the sample size is small relative to the size of the population, the sampling distribution for p is approximated closely by the binomial distribution. This is not surprising. If we were to draw a sample of 100 employees from a factory employing 10,000, the probability of selecting the first member of the sample would be $\frac{1}{10,000}$; the second member given the selection of the first, $\frac{1}{9,999}$; the third given the

¹ The formula is known as the hypergeometric, and its application to all possible values of p gives rise to the hypergeometric probability distribution.

selection of the first two, $\frac{1}{9,998}$; and so on. Thus, although the probabilities are not the same, the differences are very small. Therefore, the use of the binomial distribution which assumes equal probabilities will be only slightly in error. As a general rule, if the sample size is 10 per cent or less of the population size, the binomial distribution gives a satisfactory approximation of the sampling distribution of p from a finite population.

Example 6.6 A simple random sample of 4 invoices is taken from a file containing 100 invoices. Note that this sample size is only 4 per cent of the population size. If $\pi = .20$ is the proportion of invoices in error in the population of 100, what is the sampling distribution of p (p representing the proportion of incorrect invoices in the sample)? The exact probabilities may be derived by the application of the conditional probability rule. Thus, for example, the probability of obtaining $p = 1.00$ —i.e., 4 out of 4 invoices with errors—is

$$P(p = 1.00) = \frac{20}{100} \times \frac{19}{99} \times \frac{18}{98} \times \frac{17}{97} = .0012.$$

In contrast, the binomial probabilities may be obtained by repeated application of the binomial formula.

p	Exact Probability	Binomial Probability
.00.....	.4033	.4096
.25.....	.4191	.4096
.50.....	.1531	.1536
.75.....	.0233	.0256
1.00.....	.0012	.0016

Thus, the binomial provides a relatively good approximation of the exact probabilities in this case.

A second important characteristic of the sampling distribution of p from a finite population is that as the sample size increases, this distribution, like the binomial, may be approximated by the normal curve. As you know, the mean and standard deviation of p are needed to approximate probabilities by means of the normal curve. The expected value of p from a finite population is π , which is the same as the expected value of p from an infinite population. However, the formula for the standard deviation of p from a finite population differs slightly from that for p from an infinite population. Thus, for a finite population

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}} \sqrt{\frac{N - n}{N - 1}}.$$

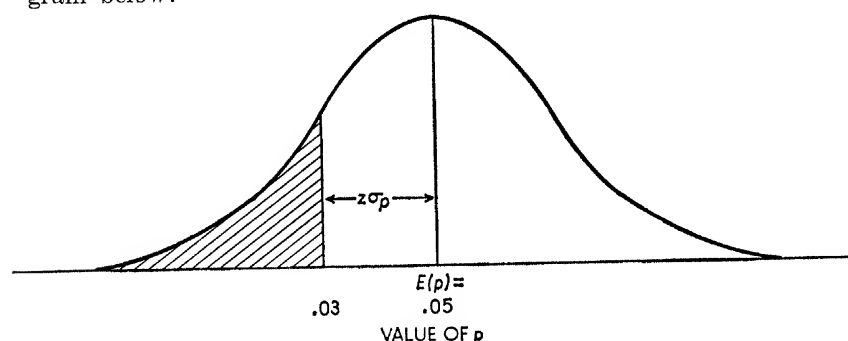
The only difference is in the factor $\sqrt{N - n/N - 1}$. Because this factor is attached to the formula for σ_p when the population is finite, it is called the *finite population correction*, or simply the f.p.c.

This "f.p.c." must be less than one. This means, when π and n are constant, σ_p will be smaller for finite populations than for infinite populations. This is reasonable. Certainly if we sample from a finite population, the amount of variability in the sample statistic should be somewhat less than if we sample from an infinite population.

Furthermore, if the sample size n is relatively small compared to the population size N —i.e., n is 10 per cent of N or less—the f.p.c. is very close to 1, and need not be used to compute σ_p . This is readily understandable for when n is small relative to N , the sampling distribution of p from a finite population is approximated by the sampling distribution of p from an infinite population.

Example 6.7 A life insurance company packs its rejected applications in batches of 500. Before the applications are stored, a random sample of 100 applications is selected to determine whether the applications should be stored away or whether the remaining applications should be examined further to see that no requests for insurance have been denied in error. The company uses the following rule: If 3 per cent or fewer of the sample of 100 indicate errors were made, the applications are stored immediately; however, if more than 3 per cent indicate errors were made, the remaining 400 are checked further.

How much protection would this decision rule give the company against neglecting to re-examine a batch of applications that had 5 per cent of the applications incorrectly rejected? Stated statistically, the problem is to determine the probability of finding 3 per cent or fewer applications defective in a random sample of 100 when $\pi = .05$. The probability desired is measured by the shaded area in the diagram below:



To determine this area under the normal curve, we must find

$$z = \frac{p - \pi}{\sigma_p}.$$

Thus,

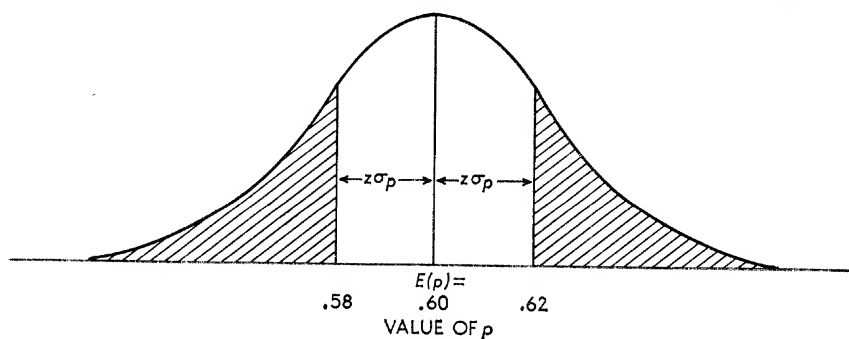
$$\begin{aligned}\sigma_p &= \sqrt{\frac{\pi(1-\pi)}{n}} \sqrt{\frac{N-n}{N-1}} \\ &= \sqrt{\frac{(.05)(.95)}{100}} \sqrt{\frac{500-100}{500-1}} \\ &= (.022)(.895) \\ &= .020.\end{aligned}$$

Substituting in the equation for z ,

$$\begin{aligned}z &= \frac{.03 - .05}{.020} \\ &= -1.00.\end{aligned}$$

When $z = 1.00$, the area in the tail is .1587. Therefore, the probability of neglecting to examine applications when 5 per cent are in error is approximately .16.

Example 6.8 A political party leader, before the party's nominating convention, conducts a survey to determine the popularity of his favorite candidate for the gubernatorial nomination. He selects a simple random sample of 400 from the half-million registered voters of his party. What is the probability that the results of the poll will be in error by 2 percentage points or more if 60 per cent of the registered voters actually do favor the candidate? An error of 2 percentage points can be made in two ways. The sample p can either be 2 percentage points less than or 2 percentage points more than π . This accounts for the two shaded areas in the following diagram:



To compute

$$z = \frac{p - \pi}{\sigma_p},$$

we first find

$$\begin{aligned}\sigma_p &= \sqrt{\frac{\pi(1-\pi)}{n}} \\ &= \sqrt{\frac{(.60)(.40)}{400}} \\ &= .0245.\end{aligned}$$

Notice that in this case we did not use the f.p.c. even though the population is finite. This is because the f.p.c. is close to 1, a fact you may verify.

When $p = .62$,

$$\begin{aligned}z &= \frac{.62 - .60}{.0245} \\ &= .82.\end{aligned}$$

When $z = .82$, the area in the tail is .2061. Therefore, the probability of getting sample proportions of .62 or greater is .2061. However, the probability of getting sample proportions of .58 or less also is .2061, because of the symmetry of the normal curve. Thus, the probability that an error of 2 percentage points or greater will be made in the survey, based on a simple random sample of 400 if $\pi = .60$, is

$$.2061 + .2061 = .4122.$$

On the basis of this result, do you consider the sampling procedure sound?

The fact that the f.p.c. is almost equal to 1 when n is small relative to N points up an important thought. Let us investigate the relationship of σ_p to N , by considering a group of 4 populations which vary in size but have the same population proportion— $\pi = .50$. The pertinent characteristics, π and N may be summarized as follows:

Population	π	N
A.....	.50	200
B.....	.50	1,000
C.....	.50	10,000
D.....	.50	Infinite

Suppose we were to select a simple random sample of 16 observations. The values of σ_p for each of the 4 populations would be as follows:

	A	B	C	D
σ_p	.1156	.1231	.1248	.1250

This is indeed an amazing result. All values of σ_p are practically equal even though the population sizes vary from 200 to infinity. Thus, for a given π , the variability of a sampling procedure and, hence, the risk of relying on such samples will be about the same even though one population is relatively small and the other relatively large.

6.6 The Sampling Distribution of the Mean

In the previous sections we were concerned with the sampling distribution of the sample proportion. The sample proportion is important when the population of interest is one of attributes and the decision-parameter is π . We turn now to a consideration of the sampling distribution of the sample mean ($\bar{x}$). The sample mean is an important statistic when the population of interest is one of variates and the decision-parameter, μ , is unknown. Thus, the material for $\bar{x}$ we will consider in this chapter is somewhat analogous to the material for p we covered earlier. However, at that time we were able to derive an *exact* expression for the sampling distribution of p with simple random sampling from an infinite population. This was called the binomial formula. In connection with $\bar{x}$, however, we shall not be able to find anything quite comparable to the binomial. However, we shall develop a normal approximation to the sampling distribution of $\bar{x}$. Fortunately, it is a very good approximation for almost all problems with which one is faced in practice.

To gain an appreciation of the reasons why this approximation is a good one, we shall begin with another simplified example. Throughout the discussion of this illustration we shall assume simple random sampling. Furthermore, the immediate example represents an enumerative study, and it therefore deals with a finite population. This means that for the time being we are going to deal with the sampling distribution of $\bar{x}$ under the conditions of simple random sampling from a finite population. Later we shall extend the results to simple random sampling from an infinite population.

Now, for the basic facts of the illustration.

Example 6.9 A firm desires to determine the average or mean age of its existing accounts receivable on the basis of a sample survey. The results of the sample are to enable the firm to decide whether additional measures are to be taken in an attempt to speed up payment of accounts. The firm decides that if they knew μ was greater than 45 days, they would take additional measures; otherwise they would take no action.

Suppose that the set of ages (in days) of the 10 accounts receivable (the small population for convenience) are, as of December 31, as follows:

<i>Name of Account</i>	<i>Age (in Days)</i>
A.....	42
B.....	28
C.....	61
D.....	30
E.....	5
F.....	15
G.....	48
H.....	25
I.....	36
J.....	40

Of course, if the firm knew this population, they would compute $\mu = 33\frac{3}{10} = 33$, and take no action. However, we are interested in what they might do if they didn't know this population and made their decision on the basis of sample information.

We will begin our analysis of the sampling distribution of $\bar{x}$ with reference to this example. Suppose, to start, we consider all possible samples of size 2 that can be drawn from the population of 10 ages. There are 45 possibilities. Of course, only one would be observed in a practical situation, but to learn about the sampling distribution of $\bar{x}$ here we are studying all of them.

Now, if our selection procedure is simple random sampling, then each one of the 45 possible samples will have equal probability of being the one selected. For each of the 45 samples, we can compute a sample mean. Each of these 45 sample means likewise has the same probability of occurrence. The distribution giving the probability of occurrence of these sample means will represent the sampling distribution of $\bar{x}$ for samples of size 2 from the population. In Table 6.1 all possible sample combinations are listed together with their associated sample means. For example, the sample mean associated with the sample based on accounts B and D is

$$\bar{x} = \frac{28 + 30}{2} = 29.$$

The next step is to transform the information in Table 6.1 into a sampling distribution of $\bar{x}$. This must show all possible values of $\bar{x}$ and their respective probabilities of occurrence. To this end, we may develop a table along the lines of Table 6.2. There, the 45 sample means have been grouped into intervals. The number of sample means that fall into each of these intervals is listed, and, more important, the probabilities of our selecting a mean in these intervals

are given. For example, there are 5 possible sample means in the interval 38–39.9 days. Since each has probability $\frac{1}{45}$ of being selected, the probability associated with that interval is $\frac{5}{45}$, or .111. In other words, the probability of selecting a sample with a mean from 38 to 39.9 days is .111 (if we were to select a simple random sample of 2 from the population under consideration).

Table 6.1

ALL POSSIBLE SAMPLE MEANS FOR SAMPLES OF SIZE 2 FROM
POPULATION OF TEN AGES OF ACCOUNTS RECEIVABLE

Accounts in Sample	Sample Variates	Sample Mean ($\bar{x}$)	Accounts in Sample	Sample Variates	Sample Mean ($\bar{x}$)
AB.....	(42, 28)	35.0	CJ.....	(61, 40)	50.5
AC.....	(42, 61)	51.5	DE.....	(30, 5)	17.5
AD.....	(42, 30)	36.0	DF.....	(30, 15)	22.5
AE.....	(42, 5)	23.5	DG.....	(30, 48)	39.0
AF.....	(42, 15)	28.5	DH.....	(30, 25)	27.5
AG.....	(42, 48)	45.0	DI.....	(30, 36)	33.0
AH.....	(42, 25)	33.5	DJ.....	(30, 40)	35.0
AI.....	(42, 36)	39.0	EF.....	(5, 15)	10.0
AJ.....	(42, 40)	41.0	EG.....	(5, 48)	26.5
BC.....	(28, 61)	44.5	EH.....	(5, 25)	15.0
BD.....	(28, 30)	29.0	EI.....	(5, 36)	20.5
BE.....	(28, 5)	16.5	EJ.....	(5, 40)	22.5
BF.....	(28, 15)	21.5	FG.....	(15, 48)	31.5
BG.....	(28, 48)	38.0	FH.....	(15, 25)	20.0
BH.....	(28, 25)	26.5	FI.....	(15, 36)	25.5
BI.....	(28, 36)	32.0	FJ.....	(15, 40)	27.5
BJ.....	(28, 40)	34.0	GH.....	(48, 25)	36.5
CD.....	(61, 30)	45.5	GI.....	(48, 36)	42.0
CE.....	(61, 5)	33.0	GJ.....	(48, 40)	44.0
CF.....	(61, 15)	38.0	HI.....	(25, 36)	30.5
CG.....	(61, 48)	54.5	HJ.....	(25, 40)	32.5
CH.....	(61, 25)	43.0	IJ.....	(36, 40)	38.0
CI.....	(61, 36)	48.5			
			Total.....1,485.0		
			Mean..... 33.0		

In short, Table 6.2 represents a sampling distribution of $\bar{x}$. A clearer mental picture of exactly what this sampling distribution of $\bar{x}$ looks like may be gained from Figure 6.3.

Remember that any sampling distribution refers to a specific statistic—this one refers to $\bar{x}$, the sample mean. It also refers to a given sample size—so far we have considered only $n = 2$. In addition, the sampling distribution in Figure 6.3 portrays the pattern of variability of $\bar{x}$ under simple random sampling from the specific population of ages of accounts receivable under study. Vary any

one of these circumstances and we would obtain a different probability distribution.

Now that we have a pretty fair idea of what the sampling distribution of $\bar{x}$ refers to, let's see what happens if we increase the size of the sample. Suppose we compute the sampling distributions

Table 6.2

**SAMPLING DISTRIBUTION OF THE MEAN FOR
SAMPLES OF SIZE 2 FROM POPULATION OF TEN AGES
OF ACCOUNTS RECEIVABLE**

Sample Mean (in Days)	Number of Possible Samples	Probability of Occurrence
10-11.9.....	1	.022
12-13.9.....	0	.000
14-15.9.....	1	.022
16-17.9.....	2	.044
18-19.9.....	0	.000
20-21.9.....	3	.067
22-23.9.....	3	.067
24-25.9.....	1	.022
26-27.9.....	4	.089
28-29.9.....	2	.044
30-31.9.....	2	.044
32-33.9.....	5	.111
34-35.9.....	3	.067
36-37.9.....	2	.044
38-39.9.....	5	.111
40-41.9.....	1	.022
42-43.9.....	2	.044
44-45.9.....	4	.089
46-47.9.....	0	.000
48-49.9.....	1	.022
50-51.9.....	2	.044
52-53.9.....	0	.000
54-55.9.....	1	.022
Total.....	45	1.000*

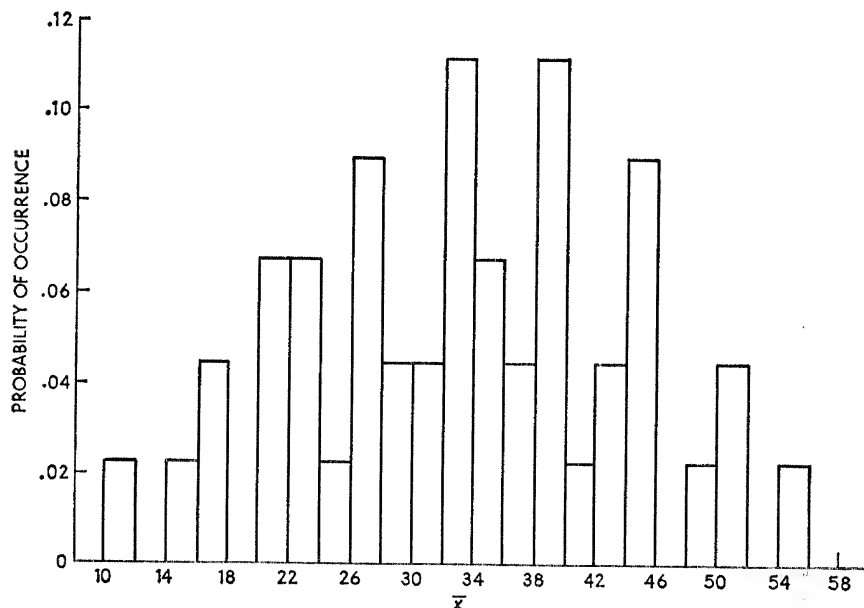
* Does not add to 1.000 because of rounding.

of $\bar{x}$ for samples of size 3, 4, 5, 6, and 7. To do this we would follow the procedure used in the last section, but Table 6.3 spares you the details. It shows us the sampling distributions for $n = 3, 4, 5, 6$, and 7 as well as for $n = 2$. Figure 6.4 portrays this analysis in terms of six simple graphs, one per sampling distribution.

One can easily notice from this chart that as the sample size increases, the sampling distribution of $\bar{x}$ becomes less variable. In this connection, Table 6.3 also shows that as the size of the sample

increases, the scatter of sample means about the population mean decreases. Thus, for $n = 2$ the probability that a sample mean falls between 30 and 35.9 is approximately .22. For $n = 3$ this probability increases to about .28. Finally, for $n = 7$ this probability rises to about .63. It increases progressively. In fact, no matter what interval about the population mean we select, the proportion of sample means within that interval increases as the sample size is increased.

Figure 6.3



One also can notice that as the sample size increases, the sampling distribution of the mean becomes more and more symmetrical. These tendencies are true not only for the sampling distribution of $\bar{x}$ from the population of 10 ages but for sampling distributions of $\bar{x}$ from almost any population.

Thus, for example, in Figure 6.5 we present a series of 4 distributions. The first figure represents a statistical population which, because it looks like a rectangle, is called a rectangular population. What might happen if we selected a simple random sample from this population? Well, if the sample size is 2, we can deduce (mathematically) that the sampling distribution of $\bar{x}$ would look like a triangle—look at the second sketch in Figure 6.5. This represents the possible and probable variation in the sample mean—it is the

sampling distribution of $\bar{x}$ for simple random samples of 2 from a population that looks like the first sketch in Figure 6.5. Similarly, the two other curves represent the sampling distributions of $\bar{x}$ associated with a rectangular population for $n = 5$ and 30.

Notice, in particular, that this set of sampling distributions, even though they all are derived from a rectangular population, change

Table 6.3

SAMPLING DISTRIBUTIONS OF $\bar{x}$ FOR $n = 2, 3, 4, 5, 6, 7$
ALL POSSIBLE RANDOM SAMPLES FROM POPULATION OF
TEN AGES OF ACCOUNTS RECEIVABLE

Sample Mean (in Days)	Probability of Occurrence for Sample of Size					
	2	3	4	5	6	7
10-11.9.....	.022					
12-13.9.....	.000					
14-15.9.....	.022	.008				
16-17.9.....	.044	.017				
18-19.9.....	.000	.017	.014			
20-21.9.....	.067	.033	.024	.008		
22-23.9.....	.067	.058	.033	.028	.010	
24-25.9.....	.022	.058	.062	.056	.038	.017
26-27.9.....	.089	.100	.100	.087	.081	.050
28-29.9.....	.044	.058	.090	.111	.129	.117
30-31.9.....	.044	.108	.105	.139	.162	.192
32-33.9.....	.111	.075	.129	.131	.176	.258
34-35.9.....	.067	.108	.110	.139	.143	.184
36-37.9.....	.044	.075	.095	.111	.133	.117
38-39.9.....	.111	.100	.095	.091	.071	.042
40-41.9.....	.022	.033	.067	.060	.043	.025
42-43.9.....	.044	.067	.033	.036	.014	
44-45.9.....	.089	.033	.029	.004		
46-47.9.....	.000	.025	.014			
48-49.9.....	.022	.017				
50-51.9.....	.044	.008				
52-53.9.....	.000					
54-55.9.....	.022					
Total*.....	1.000	1.000	1.000	1.000	1.000	1.000

* Do not add to 1.000 in some columns because of rounding.

shape as the sample size increases—they get less variable and more and more resemble a normal curve.

Figure 6.6 presents a U-shaped population, and Figure 6.7 a population skewed to the right together with sampling distributions of $\bar{x}$ for $n = 2, 5$, and 30. Again the sampling distributions become less variable and more normal.

Finally, Figure 6.8 represents a normal population distribution. It is not surprising, therefore, that all the sampling distributions

Figure 6.4

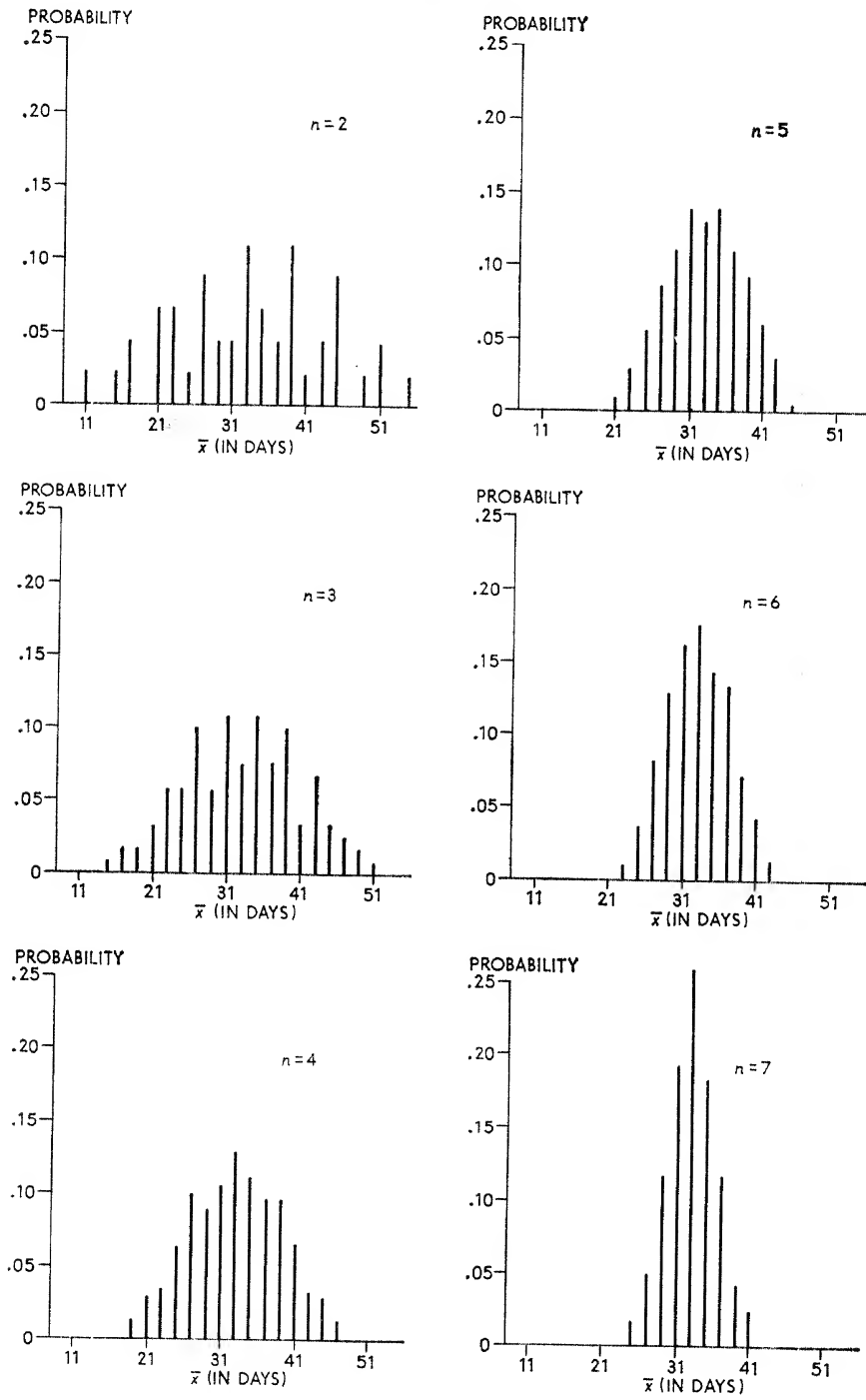


Figure 6.5

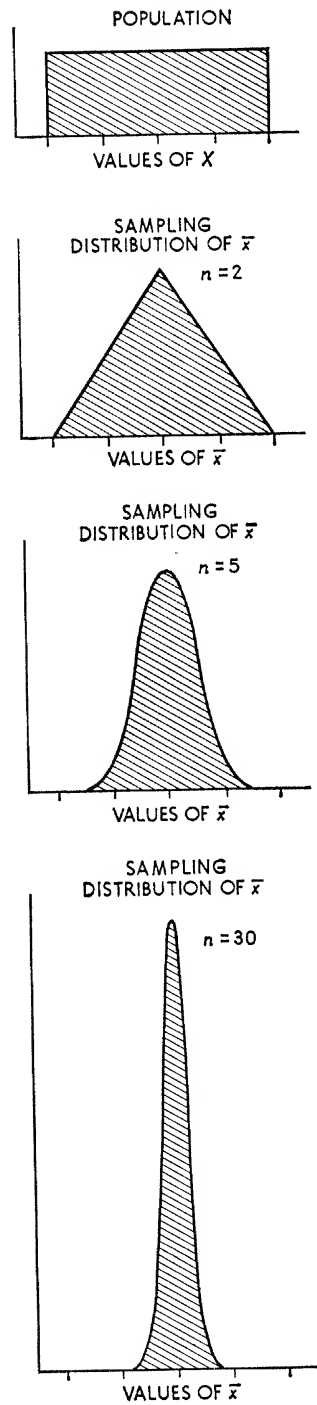


Figure 6.6

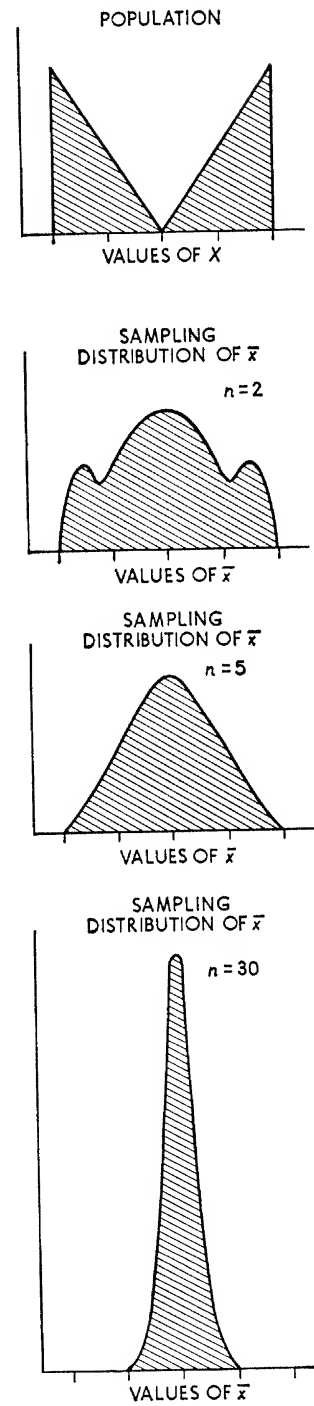


Figure 6.7

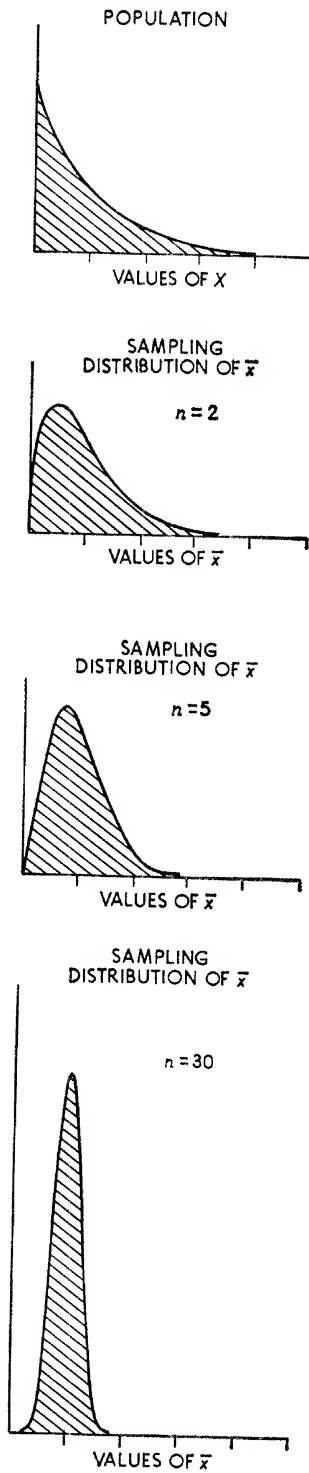
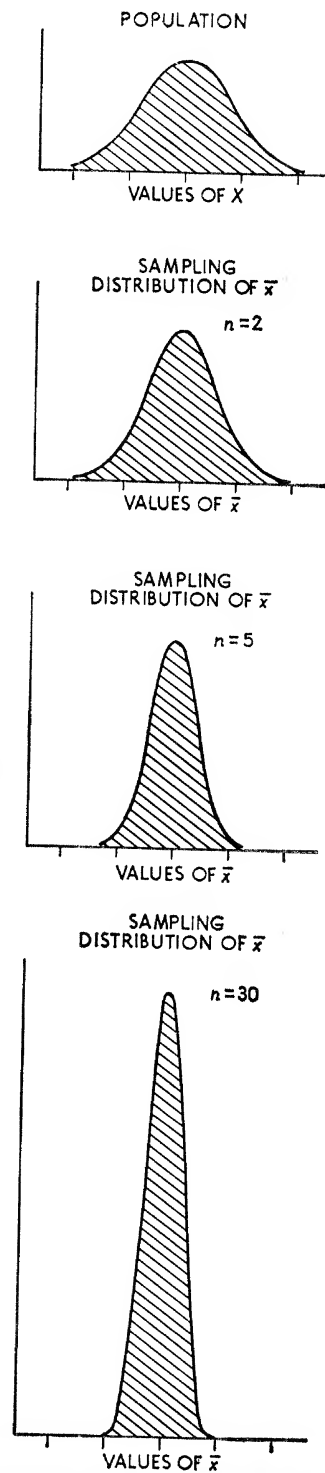


Figure 6.8



from this population also are normal. Notice, however, that as n increases the sampling distribution of $\bar{x}$ becomes less variable.

If you now compare the 4 sampling distributions for $n = 30$ in Figures 6.5 and 6.8, you will observe that they are remarkably alike even though they are derived from four different populations. This is indeed an amazing phenomenon. In fact, for most problems one meets in practice, *no matter what the shape of a population, the sampling distribution of the mean is approximated by the normal curve, if n is sufficiently large.* This means the normal distribution can be used to approximate the sampling distribution of $\bar{x}$ and, hence, to measure sampling risks. To use this result, however, we first must study the expected value and standard error of $\bar{x}$.

6.7 Characteristics of the Sampling Distribution of $\bar{x}$

As an aid in our discussion of the expected value of $\bar{x}$ and of the standard error (or standard deviation) of $\bar{x}$, suppose that we consider all possible sample means that might be drawn from a simple population. Thus, suppose we have a population of the ages of 4 typewriters—1, 2, 3, and 4 years. If 2 typewriters are to be selected at random from this population, there are 6 possible samples, namely:

Sample	$\bar{x}$
(1,2).....	1.5
(1,3).....	2.0
(1,4).....	2.5
(2,3).....	2.5
(2,4).....	3.0
(3,4).....	3.5

Each of these samples is equally likely, and, hence, so are each of the 6 values of $\bar{x}$. The sample mean is a random variable. Its expected value is easily computed as follows:

$$\begin{aligned}
 E(\bar{x}) &= \frac{1}{6} \times 1.5 + \frac{1}{6} \times 2.0 + \frac{1}{6} \times 2.5 + \frac{1}{6} \times 2.5 + \frac{1}{6} \times 3.0 + \frac{1}{6} \times 3.5 \\
 &= \frac{15}{6} \\
 &= 2.5.
 \end{aligned}$$

This refers to the average value of $\bar{x}$ we would expect to obtain if we repeated our sampling procedure over and over again.

Also notice that the mean of the population of 4 ages is

$$\mu = \frac{1 + 2 + 3 + 4}{4} = 2.5.$$

Thus, the expected value of the sample mean is equal to the mean of the population. This always will be true with simple random sampling (but not necessarily true with other types of sampling). Thus, we have

$$E(\bar{x}) = \mu$$

as a characteristic of simple random sampling. Of course, any one value of $\bar{x}$ may deviate from μ —the fact that $E(\bar{x}) = \mu$ only expresses a property of the sampling procedure.

The second characteristic of the sampling distribution of $\bar{x}$ we shall consider is a measure of its variability. This, as noted, is called the standard error of $\bar{x}$, or synonymously, the standard deviation of $\bar{x}$. This measure is symbolized by $\sigma_{\bar{x}}$. Generally speaking, it measures variability in the possible values of the sample mean. Hence, the smaller $\sigma_{\bar{x}}$, the smaller is the variability in sample means and the greater the reliability of a sampling procedure.

We now may note that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

The formula may be derived mathematically, but here we must be satisfied to check it out by a numerical example.

Reconsider the population, 1, 2, 3, and 4 years. The 6 possible values of $\bar{x}$, for samples of size 2, are

$$1.5, 2.0, 2.5, 2.5, 3.0, 3.5.$$

Their mean, $E(\bar{x})$, is 2.5. In addition, their standard deviation, $\sigma_{\bar{x}}$, is

$$\begin{aligned} \sigma_{\bar{x}} &= \sqrt{\frac{(1.5 - 2.5)^2 + (2.0 - 2.5)^2 + \dots + (3.5 - 2.5)^2}{6}} \\ &= \sqrt{\frac{1 + .25 + 0 + 0 + .25 + 1}{6}} \\ &= \sqrt{.4167} \\ &= .65. \end{aligned}$$

We should obtain the same result for $\sigma_{\bar{x}}$ by the formula we considered earlier,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

To apply this formula we first must compute σ , the standard deviation of the population. For our simple example,

$$\begin{aligned}\sigma &= \sqrt{\frac{(1 - 2.5)^2 + (2 - 2.5)^2 + (3 - 2.5)^2 + (4 - 2.5)^2}{4}} \\ &= \sqrt{1.25} \\ &= 1.12.\end{aligned}$$

In addition, $N = 4$ and $n = 2$. Thus, we have

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{1.12}{\sqrt{2}} \sqrt{\frac{4 - 2}{4 - 1}} \\ &= .65.\end{aligned}$$

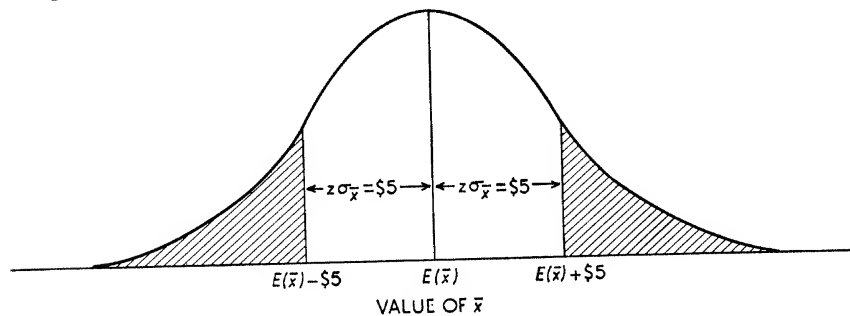
Note that the result we obtain by the use of this formula is the same as the answer derived by taking the standard deviation of all possible sample means. The use of this formula for $\sigma_{\bar{x}}$ enables us to calculate the value of the standard deviation of $\bar{x}$ without writing down all possible sample means. It is, so to speak, a short-cut formula. It also will serve to give us some insight into the meaning of $\sigma_{\bar{x}}$.

First let us see how our knowledge of the mean and the standard error of the sampling distribution of $\bar{x}$ permits us to use the normal curve to calculate risks of relying on $\bar{x}$.

Example 6.10 A manufacturer of temperature control equipment plans to take an end-of-the-year inventory of his goods in process. The goods in process are stored in 500 barrels to facilitate the work. The barrels are assigned serial numbers. He plans to select a simple random sample of 100 barrels and to determine the average value per barrel. He will multiply this average by 500 to determine the value of the inventory.

Suppose that σ , the standard deviation of the inventory for the 500 barrels, is estimated as \$50. What is the probability that the average value per barrel will be in error by \$5.00 or more?

We know that the sampling distribution of the mean is normally distributed, with $E(\bar{x}) = \mu$. The probability in question, therefore, is represented by the shaded area in the diagram below:



Note that the area in both tails of the normal curve must be considered because the sample mean may understate or overstate the population mean. To determine areas under the normal curve we first compute the normal deviate

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}.$$

Observe that in this problem we are not given the values of either $\bar{x}$ or μ . However, to determine the normal deviate we need only know the difference between the two. This difference has been specified as \$5.00.

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \\ &= \frac{50}{\sqrt{100}} \sqrt{\frac{500-100}{500-1}} \\ &= 4.48.\end{aligned}$$

Hence,

$$\begin{aligned}z &= \frac{5.00}{4.48} \\ &= 1.12.\end{aligned}$$

When $z = 1.12$, the normal curve indicates that the area in one tail is .1314. Hence, the area in both tails is .2628. In other words, the probability of an error of \$5.00 or more in the estimated average is .2628. This ordinarily would be a rather large risk to assume. For one thing you will recognize that an error of \$5.00 in the average value per barrel represents an error of \$2,500 in the aggregate value of all 500 barrels.

The formula that we have considered for $\sigma_{\bar{x}}$ refers to the sampling distribution of $\bar{x}$ from populations which are finite. You will recognize the expression $\sqrt{N-n}/\sqrt{N-1}$ as the finite population correction factor (f.p.c.) used in connection with the formula for σ_p . As N gets larger, the f.p.c. gets closer and closer to 1. Thus, for an infinite population the formula for the standard error of $\bar{x}$ becomes

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

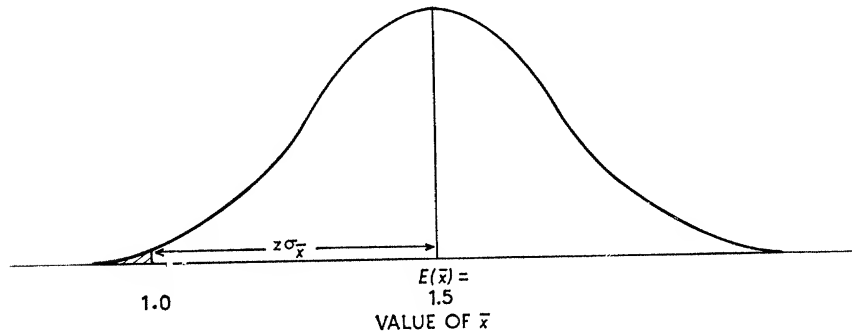
In fact, from a practical point of view, we can neglect the f.p.c. whenever the sample size is 10 per cent of the population size or less since in such instances the f.p.c. is sufficiently close to 1 so that it can be disregarded.

Example 6.11 The Department of Weights and Measures of a large city instructs its inspectors to use the following rule in testing

scales: Weigh 4 standardized weights on the scale; if the average error is more than 1 ounce, condemn the scale, otherwise approve it.

Suppose that the standard deviation of the errors is .4 of an ounce. What would be the probability of approving a scale which on the average is 1.5 ounces overweight?

The sampling distribution of $\bar{x}$ is approximated by the normal curve with $E(\bar{x}) = 1.5$ ounces. The probability in question is represented by the shaded area in the diagram below:



To determine this area we must find the normal deviate

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

But,

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{.4}{\sqrt{4}} \\ &= .2.\end{aligned}$$

Notice that we do not use the f.p.c. in this instance because the study is analytical—we are interested in the infinite population of errors associated with the weighing process.

Thus,

$$\begin{aligned}z &= \frac{1.0 - 1.5}{.2} \\ &= -2.50.\end{aligned}$$

When $z = 2.50$, the normal curve indicates that the area in the tail is .0062. Thus, the probability of approving a scale, which on the average is in error by 1.5 ounces above the true weight, is .0062 under the given decision rule.

You will observe that the formula for $\sigma_{\bar{x}}$ indicates that its value depends on n , σ , and N (when the f.p.c. is used). Let us first consider the relationship between $\sigma_{\bar{x}}$ and the sample size, n . We noticed in the last section that the sampling distribution of $\bar{x}$ became less

variable as we increased the sample size. If $\sigma_{\bar{x}}$ is to be a measure of this variability, it too should decrease with increasing n . A quick glance at the formula for $\sigma_{\bar{x}}$ will show that this is true—as n gets larger, $\sigma_{\bar{x}}$ gets smaller.

The formula indicates further that the value of $\sigma_{\bar{x}}$ varies directly with the value of the population standard deviation. In other words, the more variable the population, the greater the variability among the possible sample means for samples of a given size.

Finally, what is the relationship of $\sigma_{\bar{x}}$ to N ? Suppose that we consider a group of 4 populations which vary in size but have the same standard deviation— σ equal to 10. The pertinent characteristics, σ and N , may be summarized as follows:

Population	σ	N
A.....	10	50
B.....	10	200
C.....	10	10,000
D.....	10	Infinite

Suppose that simple random samples of size 4 were selected from each of these populations. What would $\sigma_{\bar{x}}$ equal for each of the 4 sampling distributions? We calculate the values of $\sigma_{\bar{x}}$ from the appropriate formulas. The results follow:

	Population			
	A	B	C	D
$\sigma_{\bar{x}}$	4.84	4.97	4.99	5.00

We again observe the amazing result which we saw in connection with the standard error of p . All 4 values of $\sigma_{\bar{x}}$ are approximately equal—despite the fact that the population sizes vary from 50 to infinity. Thus, it may be said that the standard error of the mean—a measure of the reliability of a sampling procedure—depends almost exclusively on the population variability (σ) and sample size (n). Population size (N) affects the variability of a sampling procedure only slightly.

6.8 Are There Sampling Distributions of Decision-Parameters?

This is a short section—only four short paragraphs. However, it deserves as much thought as any other. It answers the question: “What about the sampling distributions of decision-parameters?”

The answer is that neither π , μ , σ , nor any other decision-parameter has a sampling distribution. We repeat, decision-parameters do not have sampling distributions.

You see, when we study a statistic like p we are considering a random variable—a measure which ordinarily will vary from sample to sample. The pattern of this variation is, in fact, exactly what a sampling distribution portrays or summarizes. But in the case of π , a parameter, we are dealing with a constant—a measure which is a property of the population under study—a measure which is constant no matter what sample we select.

Thus, the answer to our question is simply—if we may repeat it once more—decision-parameters do not have sampling distributions.

6.9 You Should Now Know That

The set of all values a statistic can assume together with their probabilities of occurrence is called a sampling distribution.

Sampling distributions provide a convenient basis for computing risks.

A given sampling distribution always must be thought of as referring to a specific statistic, to a method of sampling, to a sample size, and to a specific population.

The sampling distribution of the sample proportion with simple random sampling from an infinite population can be computed exactly by the binomial formula. The sampling distribution of p depends on π and n .

The mean of the sampling distribution of p , called the expected value of p , and symbolized by $E(p)$, equals π whether a population is finite or infinite if simple random sampling is used.

Under the same conditions, the standard deviation of this sampling distribution, called the standard error of p , is

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

if the population is infinite or very large in relation to the sample size, and

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}} \sqrt{\frac{N - n}{N - 1}}$$

if the population is finite.

The sampling distribution of p becomes more and more “normal” for larger and larger values of n .

The normal curve often can be used to approximate the sampling distribution of p .

The sampling distribution of $\bar{x}$ represents all possible values $\bar{x}$ can attain, together with the probability that each of these values will occur. It represents possible values of $\bar{x}$ for samples of a given size under a certain type of probability sampling from a specific population.

The mean of the sampling distribution of $\bar{x}$, called the expected value of $\bar{x}$, equals the population mean under simple random sampling. Symbolically, $E(\bar{x}) = \mu$.

Under the same conditions the standard deviation, or standard error, of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}},$$

depending on whether the population under consideration is infinite or finite.

The sampling distribution of $\bar{x}$ generally becomes more and more "normal" as n is increased.

The normal curve often can be used to approximate the sampling distribution of $\bar{x}$.

Standard errors, such as σ_p and $\sigma_{\bar{x}}$, measure the variability of a statistic for a given sampling procedure. Hence, the smaller the standard error of a statistic, the smaller the risk in relying on sample information.

6.10 You Should Now Be Able to Solve

1. A trade association selects a simple random sample of 100 firms from a membership of 510 to determine the average hourly wage for various job classifications. Will the sampling distribution of the average hourly wage be the same for all job classifications? Explain your answer.

2. A simple random sample of 100 housewives is selected in order to estimate the proportion who favor a new cake mix. Does the binomial distribution describe the sampling distribution of the relevant statistic? Does the normal curve? Explain your answers.

3. An accountant uses the following decision rule in auditing accounts receivable: Select a simple random sample of 2,000 entries from a total of more than 100,000 entries. If 3 per cent or more of the 2,000 entries are in error, check all the entries; if less than 3 per cent are in error, approve the accuracy of the ledger.

What is the probability that the accountant will approve the accuracy of the receivables when actually 5 per cent of the entries are in error? What is the probability that he will *not* approve them on the basis of the sample if $2\frac{1}{2}$ per cent of the entries are in error?

4. Explain the following statement: The standard error of a statistic is a measure of the risk in making a decision on the basis of sample information.

5. Until recently it often was a general practice in industry to select a 10 per cent sample from an incoming lot of material to determine whether to accept or not to accept the lot. Evaluate the soundness of this procedure.

6. Unemployment, as measured by the proportion of the labor force unemployed, is known to be much higher in one state than another. The proportion of unemployed is to be estimated by a sample survey in each state. It is desired to have the same sampling error, as measured by the standard error of p , for both states. Would you or would you not suggest the same size sample for both states? Explain your answer.

7. A firm conducts a sample study to determine the average time taken to engrave calibrations on control panel dials. Twenty-five observations are to be made, and the value of $\bar{x}$ used as an estimate of μ . On the basis of this estimate, the cost of the engraving operation for an order of 10,000 dials will be estimated. If it is assumed that the standard deviation of the times is 10 minutes, what is the probability that the value of $\bar{x}$ overstates μ by 5 minutes or more? What is the probability that the value of $\bar{x}$ understates μ by 5 minutes or more? What is the probability of an error of 5 minutes or more in either direction?

8. A company manufactures cylinders that average 2 inches in diameter. The standard deviation of the diameters of the cylinders is .10 inches. The diameters of a sample of 4 cylinders are measured every hour. The sample average is used to decide whether or not the manufacturing process is operating satisfactorily. The following decision rule is applied: If the average diameter for the sample of 4 cylinders is equal to 2.15 inches or more, *or* equal to 1.85 inches or less, stop the process. If the average diameter is more than 1.85 inches and less than 2.15 inches, leave the process alone.

What is the probability of stopping the process if the process average, μ , remains at 2.00 inches?

What is the probability of stopping the process if the process

average were to shift to $\mu = 2.10$ inches? To $\mu = 1.90$ inches? To $\mu = 2.05$ inches?

What is the probability of leaving the process alone if the process average were to shift to $\mu = 2.15$ inches? To $\mu = 1.85$ inches? To $\mu = 2.30$ inches? To $\mu = 1.90$ inches?

6.11 You Will Also Find That

For review and/or further information on the sampling distribution of p and on the sampling distribution of $\bar{x}$, see:

HIRSCH, WERNER Z. *Introduction to Modern Statistics*, pp. 111-23, 140-51. New York: MacMillan Co., 1957.

LEWIS, E. E. *Methods of Statistical Analysis in Economics and Business*, pp. 195-201, 256-63, 270-78. Boston: Houghton Mifflin Co., 1953.

ROSANDER, A. C. *Elementary Principles of Statistics*, pp. 189-204, 240-46. New York: D. Van Nostrand & Co., Inc., 1951.

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics: A New Approach*, pp. 345-81. Glencoe, Ill.: The Free Press, 1956.

CHAPTER . 7

Risk and Uncertainty in Estimation Problems

7.1 Estimation Decision Problems

In Chapter 1 and at several subsequent points we have distinguished between estimation and testing problems on the basis of whether a decision problem involves numerical or nonnumerical courses of action. One reason that this distinction is important, as you will see, rests with the fact that the risks of incorrect decisions may be measured differently for the two types of problems. Furthermore, in planning a sample study, some differences in the statement of objectives also are involved, depending on whether a study involves estimation or testing.

Thus, in this chapter, we shall consider some theory, methods, and applications dealing with estimation decision problems. The next chapter will consider similar aspects of testing decision problems.

Generally speaking, a statistical estimation problem involves selecting an estimate which approximates the value of a decision-parameter. Suppose that we refresh and extend our recollection of estimation decisions by way of some illustrations.

Example 7.1 A not uncommon estimation decision problem arises whenever an individual must purchase a pane of glass to replace a broken window. Thus, one must decide on the length and on the width of the pane to be purchased. Both of these values may be estimated on the basis of observations—measurements of the window frame. Hence, you can see the problem is statistical—particularly if you recognize that repeated measurements of the length (as well as the width) would, to some extent, vary and that two or three or even one dozen measurements on the same window would represent only a

sample of measurements. Thus, the choice of a course of action must depend on a sample of observations.

Example 7.2 An interesting estimation decision problem in the field of accounting arises in the settlement of inter-airline accounts.¹

The problem arises because of situations such as the following. Suppose a passenger wishes to travel from Portland, Oregon, to Seattle, Chicago, and New York City. If he purchases a complete ticket in Portland, United Air Lines will collect the entire fare (\$158.85) for the travel indicated. However, UAL will carry the passenger only from Portland to Seattle. Northwest Air Lines would provide transportation from Seattle to Chicago, and Trans-World Air Lines from Chicago to New York.

Thus, the problem centers on estimating the amount of the total fares UAL owes NWA and TWA—not for one ticket but perhaps for 20,000 such tickets sold during a given month. Alternatives are represented by the various estimates of revenue to be paid to NWA and TWA. United Air Lines, and other companies as well, have developed statistical decision rules based on sample information to decide upon the choice of alternatives and, hence, upon the settlement of interline accounts.

In the two preceding illustrations, a course of action would follow directly from an estimate of the decision-parameter. In many other business decisions, however, it is practically impossible to formulate one's problem so that action will flow from such estimates alone. Other factors, some qualitative, must be taken into account together with estimates of several decision-parameters in order to decide on a course of action.

Example 7.3 One management textbook² lists the following as criteria which should be considered in the location of a business: (a) proximity to desired markets, (b) proximity to supply of raw materials, (c) availability of labor, (d) adequacy and cost of transportation, (e) adequacy and cost of power and fuel, (f) proximity to firms in similar fields, (g) suitability of climate, (h) character of the community and special inducements, (i) reasonableness of state and local taxes, and (j) suitability to decentralization.

Some of these criteria, such as the character of the community, must be evaluated qualitatively in making a decision on where to locate. However, others, such as the average temperature, rainfall, and other weather factors bearing on the suitability of climate, may be evaluated quantitatively and estimates of these quantities derived

¹ Winston C. Dalleck, "Inductive Accounting: Settling Interline Accounts by Sampling Methods," *Industrial Quality Control*, Vol. XIII, No. 6 (December, 1956), pp. 12-16.

² John A. Shubin, *Business Management: Efficiency Surveys and Systems* (New York: Barnes & Noble, Inc., 1954).

on the basis of statistical studies. Thus, some of the bases for decision are provided by estimates of decision-parameters.

Problems of this type are not wholly statistical decision problems. They merely contain some statistical aspects. The statistical part of such a problem is to estimate pertinent decision-parameters, knowledge of which would contribute to the basis of a managerial decision. Of course, reliable estimates are important even though the decision problem is not wholly statistical.

Thus, we see that statistical estimation is important in some problems as a direct basis for decision; and in others as a means of contributing to a basis for decision. Often, of course, the way in which a problem is formulated governs the way in which such estimates will be used—directly or indirectly. In any event, estimates must be made to serve a useful function if they are to be worth their cost. Hence, one always must plan, in a general way, how he will use them. In addition, to further motivate our study of estimation, we may note that unless estimates are computed by reliable methods, they will be unable to serve a useful function.

The above illustrations imply the necessity of primary studies—studies geared to the particular problem at hand. It also may be noted that many secondary studies involve estimation. That is, from the point of view of the producer of such data, the purpose of his studies is to provide information for the use of consumers of the data. This information ordinarily may be provided most conveniently in the form of estimates of proportions or averages or aggregates or whatever else is of interest.

Example 7.4 A few years ago *Life* magazine sponsored an extensive and comprehensive statistical study of consumers' buying habits in the American market. The study, based on a probability sample, was conducted as a public service to provide quantitative information as a background for the marketing decisions of individual businesses. The survey was conducted for *Life* under the auspices of the Alfred Politz Research, Inc.

Among the hundreds of estimates prepared as a result of the study were estimates of *average* annual household expenditures for various goods and services by families with different annual incomes, estimates of the *proportions* of income spent by families on different goods and services, estimates of the *frequency distribution* of households by annual income, etc. Thus, information of interest to those who make marketing decisions was presented clearly and concisely in terms of estimates of decision-parameters.

As has been already noted, estimation problems involve selecting, on the basis of sample information, estimates which approxi-

mate the values of decision-parameters. Of course, we want, in these problems, to learn how to derive good approximations—that is, to use sample information most efficiently. Thus, our main task in this chapter is to learn what rules one should follow in this endeavor. For example, it may seem reasonable to use the value of a sample statistic (such as $\bar{x}$) as an approximation to a population parameter (such as μ)—but this may be a risky business. The question that arises, therefore, is: “How can we appraise this decision procedure?”

In discussing estimation it is customary to distinguish two types of estimates—generally called (a) point estimates, and (b) interval estimates. The difference between (a) and (b) is quite simple and may be pointed up by considering an example.

Example 7.5 A recent sample survey of assessed valuations in the United States reports estimates of the per cent of total assessed value that consists of “residential one-family houses.” The estimates are reported both as point estimates and interval estimates. Thus, for the state of Virginia estimates are as follows:

Type of Estimate	Per Cent of Total Assessed Value
Point.....	47.6
Interval.....	46.3 to 48.9

As you can see, the point estimate is a single value which attempts to “hit” the true but unknown population ratio. On the other hand, the interval estimate is a range of values which it is hoped includes the true but unknown population ratio.

As this illustration implies, a point estimate is a single estimate of the value of a decision-parameter, while an interval estimate takes the form of a range of estimates. Making a point estimate is like throwing a dart at some point on a game board, while making an interval estimate is analogous to tossing a hoop at the point and trying to encircle it. Of course, whether one throws a dart or a hoop depends on the game he is playing. Similarly, whether one makes a point estimate or an interval estimate depends on his decision problem. For the present we shall center attention on point estimation. In Section 7.6 we shall consider procedures for constructing interval estimates.

Let's consider one other introductory thought about estimation decision problems. This involves an answer to the question: “What form does a statistical decision rule assume in such a problem?” As we know, a statistical decision rule relates actions to possible observations before a study is carried out. It is a complete plan on how to sample and on how to use the sample obtained. Furthermore, it is

important to remember it is formulated a priori. These are general ideas. To begin being more specific about the form of such a rule in an estimation problem, suppose we consider another illustration.

Example 7.6 A plastics firm must decide on a piecework wage rate for a buffing and polishing operation. The average time for this operation is an important factor in determining the rate. Since the population of observations is infinite because the study is analytical, a sample study is planned to estimate the mean time for the operation. The choice of a wage rate will be based on this estimate. Hence, the time-and-motion-study man suggests the following statistical decision rule: Select a simple random sample of 20 and use the value of the sample mean, $\bar{x}$, to estimate the population mean, μ . The procedure indicated obviously involves statistical estimation.

As this example implies, there are at least three integral parts to a statistical decision rule in an estimation problem.

One is the instruction on what method of sample selection to use; e.g., simple random sampling.

The second is the instruction on how large a sample to select; e.g., 20 observations on the operation.

The third is a rule or formula for translating the sample observations into an estimate of the decision-parameter of interest; e.g., use the value of $\bar{x}$ as an estimate of μ . This rule or formula is called an *estimator*.

All three of these integral parts of a decision rule should be formulated in the planning stage of a study. However, they should not be formulated haphazardly. There are certain principles that must be followed. In this chapter, we shall consider the problem of formulating decision rules for estimation problems and how a decision maker can rationally answer such questions as: "What formulas (estimators) should we use to construct estimates?" "When should simple random sampling be preferred to other types of probability sampling?" "How do we determine sample size?"

7.2 Estimators

As was indicated in the last section, one integral part of a statistical decision rule in an estimation problem is the estimator—the formula which translates sample observations into an estimate of a decision-parameter.

Among the simplest and most reasonable estimators are: (1) $\bar{x}$, the sample mean, as an estimator for μ , the population mean, and (2) p , the sample proportion as an estimator for π , the population proportion. Although these estimators undoubtedly seem reason-

able to you, it probably is difficult for you to explain why they do. What does "reasonable" mean in this connection? Think it over.

In any event, our present task is to give "reasonableness" a more precise statistical meaning. We must consider certain criteria for evaluating estimators—for sorting out reasonable estimators from unreasonable estimators.

Note, first of all, that we cannot always select the corresponding statistic as an estimator for a decision-parameter. For example, the sample aggregate would be a poor estimator for a population aggregate. If you don't think so, consider the following problem: Suppose you wanted to estimate the total amount of change in the pockets of all your fellow students (e.g., at next Monday's class). Would the aggregate in a sample of two pockets be a reasonable estimate of the aggregate amount of change in everyone's pockets? You may reason out that a better estimator for a population aggregate is $N\bar{x}$. Why?

In our discussion of criteria for evaluating estimators the following notation should be kept in mind. First of all, π , μ , A , etc., as always, will represent decision-parameters—population properties, the values of which we desire to estimate. Secondly, an estimate of any decision-parameter will be symbolized by a caret over π or μ or A , etc. That is, estimates will be symbolized by $\hat{\pi}$, $\hat{\mu}$, $\hat{A}$, etc.

Thus, when we express any estimator formula-wise, we should write, for example,

$$\hat{\pi} = p,$$

which says "estimate the value of π by using the value of the sample proportion." Similarly, other estimators we already have discussed are, in this notation,

$$\hat{\mu} = \bar{x}$$

and

$$\hat{A} = N\bar{x}.$$

Now, with this notation in mind, let's consider three criteria which statisticians often use to evaluate the reasonableness of estimators. These criteria or properties are called *consistency*, *unbiasedness*, and *efficiency*.

Roughly speaking, an estimator is *consistent* if, as the sample size is increased, the probability that an estimate will be approximately equal to the decision-parameter tends closer and closer to one. To be more precise—in the case of a finite population—we may

say an estimator is consistent if an estimate becomes exactly equal to the decision-parameter when $n = N$.

Example 7.7 A mailing list service sells lists of names for use in advertising and promotional campaigns. Some lists include persons in specified income brackets, others include only doctors or only lawyers, and still others include persons in given geographical areas. Suppose that the service must decide on a price to charge for two such lists. To do so they must know the number of persons common to both lists.

Assume that one list contains N names, the other M names, and simple random samples of n and m names, respectively, are selected from these lists. If there are d duplicates between the samples, we may estimate the number of duplicates, D , in both complete lists by the estimator

$$\hat{D} = \frac{NM}{nm} d.$$

This estimator has the property of being consistent. Note that if $n = N$ and $m = M$, d must equal D because a complete count is being taken. In such a case our estimator becomes $\hat{D} = D$. Any estimate is exactly equal to the decision-parameter, D .

For an illustration of the use of this estimator, suppose that we have two lists with 10,000 and 5,000 names, respectively. If a sample of 400 names is selected from each list and 7 duplicates observed, we would estimate D as

$$\begin{aligned}\hat{D} &= \frac{(10,000)(5,000)}{(400)(400)} (7) \\ &= 2,188.\end{aligned}$$

Consistency, or the lack of it, may be interpreted in terms of a sampling distribution. Thus, we know that the sampling distribution of $\bar{x}$ and the sampling distribution of p tend to center more and more closely about μ and π , respectively, as n is increased. Hence, the estimators $\hat{\pi} = p$ and $\hat{\mu} = \bar{x}$ are consistent. So is the estimator $\hat{A} = N\bar{x}$ a consistent estimator.

Consistency is a desirable property. If an estimator is consistent, this means that the larger the sample, the greater the reliability of one's estimate. However, you will note that the criterion of consistency refers to a property of the sampling procedure. It refers to what results an estimator would yield if the sample size were increased. Of course, the consistency of an estimator does not guarantee a good estimate in a single sample. Consistency is not a property of any one estimate. It simply is a property of an estimating procedure. This also is true of all other criteria for estimators. Re-

member no one can determine whether or not a single sample or a single estimate will be good or not.

An estimator is *unbiased* if its expected value equals the decision-parameter being estimated. For example, $\hat{\pi} = p$ is unbiased if p is the result of simple random sampling. Why? With simple random sampling, $E(p) = \pi$. Similarly, with simple random sampling, $\hat{\mu} = \bar{x}$ and $A = N\bar{x}$ are unbiased because

$$E(\bar{x}) = \mu, \text{ and } E(N\bar{x}) = N\mu = A.$$

Example 7.8 One simple example of an estimator which is biased is

$$\hat{\sigma}^2 = s^2.$$

That is, the sample variance is *not* an unbiased estimator of the population variance. The expected value of s^2 does not equal σ^2 . Instead it can be demonstrated mathematically that for a finite population

$$E(s^2) = \frac{N}{N-1} \cdot \frac{n-1}{n} \sigma^2.$$

For an infinite population this becomes

$$E(s^2) = \frac{n-1}{n} \sigma^2.$$

(These formulas indicate that on the average the value of a sample variance is smaller than that of its population's variance. Why?)

Statisticians often prefer

$$\hat{\sigma}^2 = \frac{N-1}{N} \cdot \frac{n}{n-1} s^2$$

as an estimator for the variance of a finite population and

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2$$

as an estimator for the variance of an infinite population. The expected values of these estimators equal σ^2 , and, hence, they are unbiased.

As a numerical example, suppose that we were to select a simple random sample of 4 ball bearings from a lot of 10,000 and found the variance of their diameters, s^2 , to be equal to .0012. If we ignore the factor $(N-1)/N$, which is close to 1, an estimate of the population variance is

$$\begin{aligned} \hat{\sigma}^2 &= \frac{4}{3} (.0012) \\ &= .0016. \end{aligned}$$

Example 7.9 In problems which deal with the statistical control of production processes, the sample range often is used to estimate the population (process) standard deviation. The great advantage of using the range is the ease and simplicity of its calculation.

For samples of a given size which are drawn from a normal population, statisticians have determined the relationship between the expected value of the sample range and the standard deviation of the population. This relationship which is designated by the symbol, d_2 , may be written as:

$$d_2 = \frac{E(R)}{\sigma}$$

or, alternately,

$$\frac{E(R)}{d_2} = \sigma.$$

$E(R)$ is the expected value of the sample range, and d_2 is a constant which depends on the sample size. Some values of d_2 follow:

Sample Size	d_2
2.....	1.128
3.....	1.693
4.....	2.059
5.....	2.326

In order to estimate σ , we, therefore, may use the range of one sample. However, rather than use the range of a single sample, the average range for a group of samples of the same size usually is used to estimate σ . That is, the unbiased estimator

$$\hat{\sigma} = \frac{\bar{R}}{d_2}$$

is used. Here, $\bar{R}$ is the average of several sample ranges.

Consider this numerical example of the use of the estimator. Suppose that we have 5 samples, each a sample of 4 diameters of steel rods. The ranges for the 5 samples are (in inches) .011, .003, .009, .023, .014.

$$\begin{aligned} R &= \frac{.011 + .003 + .009 + .023 + .014}{5} \\ &= .012 \text{ inches.} \end{aligned}$$

For samples of size 4, $d_2 = 2.059$. Therefore, an estimate of the population standard deviation is

$$\begin{aligned} \hat{\sigma} &= \frac{.012}{2.059} \\ &= .0058 \text{ inches.} \end{aligned}$$

Thus, unbiasedness is a long-run property of an estimator. It means that if the estimating procedure were repeated over and over again many times, on the average the estimator would yield the right answer. Thus, with simple random sampling, $\bar{x}$ is an unbiased estimator of μ because if we used $\bar{x}$ over and over again to estimate μ , on the average our estimate would exactly equal μ .

Unbiased estimators seem reasonable, if for no other reason

than that they appeal to one's sense of fairness. However, it must be remembered that unbiasedness simply means that the estimating process is headed in the right direction. Any one estimate, of course, may be in error. Therefore, we also must consider the magnitude and probabilities of possible errors in evaluating estimators. We must consider the potential variability of an estimating procedure. This leads to the criterion of *efficiency*.

As you know we may measure the variability in a sampling procedure and, hence, in an estimating procedure by the standard error or by the variance of a statistic. For example, the variance of $\bar{x}$ for simple random sampling from a finite population is

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \frac{N-n}{N-1}.$$

This measures the variability in the procedure of using $\bar{x}$ as an estimator of μ .

Efficiency is a relative criterion for evaluating estimators. That is, it involves the comparison of two estimators by comparing their variances. Thus, one estimator is more efficient than another if its variance is smaller. Note, then, that more efficient means *more precise* and *less variable*.

Example 7.10 Suppose that we wish to estimate the arithmetic mean of a population that may be assumed to be normally distributed. One estimator is, of course,

$$\hat{\mu} = \bar{x},$$

which we know to be unbiased (for any type of population). In addition, in a normal population the mean and the median are equal. Thus we might use the sample median as an estimator for μ . Fortunately, this estimator also is unbiased and at times quite useful.

However, the sample mean is a more efficient estimator of μ than the sample median; that is, it has a smaller standard error. As we know, the standard error of $\bar{x}$ is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(for an infinite population). In contrast, the standard error of the sample median, for large samples from a normal population, is equal to

$$1.2533\sigma_{\bar{x}}$$

and, hence, is obviously larger than $\sigma_{\bar{x}}$.

Example 7.11 Unbiased estimators are not necessarily preferable to biased estimators. At times the latter are more efficient. For example, a commonly used, though biased procedure, is called *ratio esti-*

mation. To illustrate this method in a simple situation assume that we wish to estimate the average retail sales in 1959 for 3 stores. Suppose that this population includes the following observations:

Store:	A	B	C
1959 Sales:	\$2,000	\$4,000	\$12,000

Hence, $\mu = \$6,000$.

If a simple random sample of 2 stores is to be selected, there are 3 possible samples and 3 possible estimates of μ —

Sample	Estimate of μ
A,B.....	\$3,000
A,C.....	7,000
B,C.....	8,000

The average (expected value) of these estimates is \$6,000, which once again demonstrates that the mean of a simple random sample is an unbiased estimator of μ . However, you will note that the possible estimates are rather variable—the standard error, in fact, is \$2,082.

In contrast to simple random sampling, we might estimate retail sales for 1959 by a ratio estimate. Thus, suppose that retail sales for 1958 also were available as follows:

Store:	A	B	C
1958 Sales:	\$1,000	\$4,000	\$10,000

This information can be used in estimating sales for 1959. Thus, suppose that a sample of 2 stores, A and C for example, were selected. The average sales for these stores in 1958 were \$5,500; those in 1959 were \$7,000. The ratio of average sales in 1959 to those in 1958 for the sample is

$$\frac{7,000}{5,500} = 1.2727.$$

If we apply this ratio to the average sales in all stores in 1958, we would obtain our ratio estimate

$$(1.2727)(\$5,000) = \$6,364.$$

Again considering all possible samples of size 2, we observe that there are 3 possible ratio estimates, namely,

Sample	Ratio Estimate of μ
A,B.....	\$6,000
A,C.....	6,364
B,C.....	5,714

Hence, the average (expected value) of these estimates is \$6,026 which is not equal to the population mean. The ratio estimating procedure is biased—though the bias is slight. However, note that the possible estimates are relatively stable—more so than with the unbiased estimator, $\bar{x}$.

Ratio estimates often are used in marketing research and economic

surveys, and they are often preferable to simple estimators despite the possibility of some bias because they often are far more efficient. Of course, auxiliary information (e.g., 1958 sales) must be available to use them.

Several other criteria besides those of consistency, unbiasedness, and efficiency have been suggested by statisticians for deciding upon estimators. One, which reflects rather recent and important developments in modern statistical theory, is called the *minimax* principle.

Example 7.12 A good part of the modern developments in the field of statistics centers on the matter of how to consider consequences more explicitly in a decision problem. For example, we might assume that the loss resulting from a poor estimate of π is proportional to the square of the difference between π and our estimate; that is, proportional to

$$(\hat{\pi} - \pi)^2.$$

Thus, if we were to estimate π as .5 when it is .4, our loss would be proportional to .01; while if we were to estimate π as .9 when it was .6, our loss would be proportional to .09.

For different estimates (our alternatives) and for different values of π (states of the world) there are different losses. Suppose, for further numerical examples, that we were sampling from a population of 3 dwelling units which may be occupied or not occupied. The population proportion of occupied dwelling units, π , might equal 0, $\frac{1}{3}$, $\frac{2}{3}$, or 1. Obviously, among our possible alternatives, estimates of π , are 0, $\frac{1}{3}$, $\frac{2}{3}$, and 1. However, we need not limit ourselves to these estimates—we may also consider $\frac{1}{6}$, $\frac{1}{2}$, $\frac{5}{6}$, and any other values as alternatives. The following table presents the consequences as measured by the loss function

$$(\hat{\pi} - \pi)^2$$

for various states of the world and different possible estimates:

Alternative Estimate	State of the World			
	$\pi = 0$	$\pi = \frac{1}{3}$	$\pi = \frac{2}{3}$	$\pi = 1$
$\hat{\pi} = 0$	.000	.111	.444	1.000
$\hat{\pi} = \frac{1}{6}$	.028	.028	.250	.694
$\hat{\pi} = \frac{1}{3}$	.111	.000	.111	.444
$\hat{\pi} = \frac{1}{2}$	.250	.028	.028	.250
$\hat{\pi} = \frac{2}{3}$	.444	.111	.000	.111
$\hat{\pi} = \frac{5}{6}$	.694	.250	.028	.028
$\hat{\pi} = 1$	1.000	.444	.111	.000

Now note that if we are to choose an alternative (estimate) on the basis of a random sample, the consequences also will depend on chance. We then may compute the expected loss of our estimating

procedure. Thus, suppose that we were to select a sample of 2 from the population of 3 dwelling units and use the estimator $\hat{\pi} = p$ as a basis for the choice of an estimate. There would be 3 possible samples of size 2 that we may draw from our population of 3 dwelling units. If $\pi = 0$, the proportion of occupied dwelling units, p , in each of the 3 possible samples of 2 is obviously 0. Hence, the probability that $\hat{\pi} = 0$ is 1, and our expected loss is 0. However, if $\pi = \frac{1}{3}$, one of the possible samples of 2 would have no occupied dwelling units or $p = 0$; 2 of the samples would have 1 occupied dwelling unit, or $p = \frac{1}{2}$. Hence, the probability that

$$\hat{\pi} = p = .00 \text{ is } \frac{1}{3},$$

while the probability that

$$\hat{\pi} = p = \frac{1}{2} \text{ is } \frac{2}{3}.$$

The expected loss of this estimator is, thus,

$$.111 \times \frac{1}{3} + .028 \times \frac{2}{3} = .056.$$

Similarly calculations show that for the estimator $\hat{\pi} = p$, the expected losses are as follows:

STATE OF THE WORLD			
$\pi = 0$	$\pi = \frac{1}{3}$	$\pi = \frac{2}{3}$	$\pi = 1$
.000	.056	.056	.000

By what criterion should we choose an estimator? One answer involves considering the maximum expected loss over all states of the world (e.g., the maximum for $\hat{\pi} = p$ is .056), and selecting that estimator which minimizes this maximum loss. Such an estimator may be derived mathematically and is called a *minimax* estimator of π because it minimizes one's maximum loss.

With simple random sampling from a finite population it may be shown mathematically that this estimator is

$$\hat{\pi} = \frac{\sqrt{np} + \frac{1}{2}\sqrt{(N-n)/(N-1)}}{\sqrt{n} + \sqrt{(N-n)/(N-1)}},$$

while for an infinite population it is

$$\hat{\pi} = \frac{\sqrt{np} + \frac{1}{2}}{\sqrt{n} + 1}.$$

Applying the estimator for a finite population to a situation for which $N = 3$ and $n = 2$, we would note that if $p = 0$,

$$\hat{\pi} = \frac{\sqrt{2}(0) + \frac{1}{2}\sqrt{\frac{1}{2}}}{\sqrt{2} + \sqrt{\frac{1}{2}}} = \frac{1}{6},$$

while if $p = \frac{1}{2}$ our estimate would be $\hat{\pi} = \frac{1}{2}$, and if $p = 1$ our estimate would be $\frac{5}{6}$. You might verify that for each state of the world the expected loss of this estimator is only .028. Hence, the maximum

loss for the minimax estimator is .028, while the maximum loss for the unbiased estimator is .056. In fact, if we were to compute the maximum loss for any other estimator of π , we still would find that the maximum loss for the minimax estimator is smaller—hence its name, minimax estimator.

Unfortunately, the use of a minimax estimator or any other similar type of estimator, at the present state of our knowledge, is limited because of the difficulty of measuring consequences explicitly in most decision problems.

Our brief discussion of estimators in this section has indicated some of the criteria statisticians use to decide between reasonable and unreasonable estimators. We considered such criteria as consistency, unbiasedness, efficiency, and the minimax principle. Although there is far more that might be said about these matters, we have covered enough ground to realize that the use of such estimators as $\hat{\pi} = p$, $\hat{\mu} = \bar{x}$, and $\hat{A} = N\bar{x}$ is far from arbitrary.

7.3 Choice of a Method of Sampling

In formulating a decision rule for an estimation problem, one must decide on a method of probability sampling at the same time the choice of an estimator is being made. In fact, an estimator such as $\hat{\mu} = \bar{x}$ has no meaning unless the method of selecting the sample, upon which $\bar{x}$ is to be based, is specified. In the preceding section we considered only simple random sampling as a basis for the estimators discussed. Now let us consider other methods of sampling as well, and an answer to the question, "For a given study, how does one decide on the method of probability sampling to be used?" For example, how should one choose from among simple random sampling, cluster sampling, stratified sampling, and sequential sampling in a marketing research study?

Generally speaking, our aim in any estimation decision problem is to choose the method of sample selection which will yield the smallest risks for a given cost or which will cost the least to achieve a given set of risks.

We know that any method of sampling is characterized by some possible variation in a sample result—and that a useful measure of this variability (when the sampling distribution is normal) is the standard error. Hence, in a given situation to compare methods of sample selection, we may compare the standard errors that are associated with those methods and select the one with the smallest standard error.

In order to discuss this topic more completely we would need to

use formulas for standard errors of $\bar{x}$ and of p , etc., for several methods of sample selection. However, we have considered such standard error formulas only for simple random sampling. Thus, in this part of our study we must rely on specific numerical examples to indicate general points of interest.

Example 7.13 During World War II, the administration of tire rationing by the government required current and reliable information on inventories held by dealers, distributors, and manufacturers.³ For reasons of economy and expeditiousness this information was almost always determined by a sample study. That is, a sample of dealers, distributors, and manufacturers was selected, and the country's inventory estimated on that basis.

The particular method of sampling often used in these studies was stratified sampling. For example, let us consider the March, 1945, survey. Then dealers were classified into classes or strata according to the number of tires they held in September, 1944, a date for which complete information was available. Thus, one group included dealers who held from 1 to 9 passenger tires on September, 1944; another included dealers who held 10–19 passenger tires; etc. A simple random sample was drawn from each stratum, and an estimate of the aggregate inventory prepared (by summing the estimates of the aggregates for each of the strata).

Stratification increased the efficiency of the sampling plan to a great extent. It was estimated that if a simple random sample had been selected, $2\frac{1}{2}$ times as many dealers would have been necessary in order to achieve the same standard error as was achieved with the stratified sample.

This was true because the strata into which dealers were grouped were fairly homogeneous with regard to the characteristic under consideration. In addition, the number of tires per dealer in one stratum differed greatly from the number of tires per dealer in the other strata. In general, the aim in stratification is to select strata which have little variability within each stratum, but for which the variability among the strata is large. When these aims are fulfilled, stratified sampling involves less risk of error than simple random sampling for the same size sample.

Stratified sampling, however, will as a rule be more costly than simple random sampling. In the case of stratified sampling there must be information which will enable us to divide the over-all frame into lists for each stratum. In the tire inventory study, data

³ W. Edwards Deming and Willard A. Simmons, "On the Design of a Sample for Dealer's Inventories," *Journal of the American Statistical Association*, Vol. XLI (1946), pp. 16–33.

on tire inventories for September, 1944, were available and, hence, the cost of constructing lists for each stratum could be kept relatively small. Cost factors must be considered, for what we desire to attain is increased precision per dollar spent. We must be sure, therefore, that added gains in precision are not whittled away by very large increases in cost.

Example 7.14 An automobile parts manufacturer, in order to make certain pricing decisions, needs information about the number of competing brands carried in inventory by gasoline stations who also sell the manufacturer's products. A marketing research firm hired to conduct the study considers two possible methods of sample selection. One is the selection of a simple random sample; the other is cluster sampling.

Thus, on one hand, a sample of gasoline stations might be drawn one at a time and at random from a list of the firm's customers. On the other hand, clusters of 5 neighboring stations might be formed and a sample of these clusters drawn at random.

The main advantage of cluster sampling is a reduction in the costs of travel which is necessary to visit stations in order to obtain information. Thus, for this study it is determined that 60 clusters of 5 stations can be selected for the same cost as a simple random sample of 100 stations.

However, it is recognized that this larger sample size may not mean greater precision in the estimates of the decision-parameters of interest. Neighboring stations undoubtedly will be similar with respect to the characteristic of interest. Hence, a cluster of 5 stations cannot be expected to provide as much different information about a population as 5 separately selected elementary units. Although we shall not compare their formulas, it may be noted that this phenomenon will reflect itself in the standard errors of the estimates. Thus, suppose that the marketing research firm estimates that for a simple random sample of 100 gas stations the variance of the proportion of firms which do not carry competing products would be about .0016. However, they also estimate that for a cluster sample of 300 gas stations the variance of p would be approximately .0012. Note that while the cluster sample does not cut the variance of p to one third of its value for simple random sampling, it does result in a somewhat smaller variance, and, hence, a gain in precision for the same cost is to be expected. The survey should be carried out by that method.

You will notice that to compare simple random sampling and cluster sampling in the above illustration, we borrowed the notion of efficiency from the last section; that is, we compared the variances associated with both methods of sampling. Thus, efficiency is a criterion for deciding (a) between estimators given a method of sampling (as in Examples 7.10 and 7.11 of the last section), and

(b) between methods of sampling (as in Examples 7.13 and 7.14 above).

Our short survey in this section indicates that the choice of a method of sampling is important in the design of a statistical study because some methods may yield more reliable and, hence, less risky results than other methods. Our discussion of the choice of a method of sampling, although quite brief, has indicated the basic principles (cost versus precision) by which such choices are made in practice. Subsequently, however, we again shall center our studies on simple random sampling.

7.4 Determining Sample Size

Another integral part of a statistical decision rule in an estimation problem is the sample size. How does one decide how large a sample to select in such a problem? The answer to this question is: "On the basis of how much sampling risk it is economical to tolerate." Let's discuss this statement further.

We already are well aware that the possibility of error in a sample result always exists. Thus, there always is the possibility of incorrect decisions in estimation problems. What are these incorrect decisions? Generally speaking there are two types: overestimating and underestimating. Of course, some estimates are poorer than others. Still generally speaking, the larger the error in any estimate of a decision-parameter, the poorer is our choice—the more serious our incorrect decision. Thus, suppose that we estimate the inventory of a product at 50,000 units and estimate its value accordingly. This is an incorrect decision if 55,000 units is the actual inventory. Yet it is not *as* incorrect as if the estimate were 40,000.

However, you also should recognize that in many problems some error or risk can be or should be tolerated. In such cases one need not aim at perfect precision—it is unnecessary. One should consider instead (1) how large an error in the estimate can be tolerated, and (2) how great a risk of exceeding that error should be taken. These two considerations are necessary before proceeding to solve a problem. They represent the way in which managerial objectives should be set in an estimation decision problem. As we shall soon see, once they are set, we may determine such things as the necessary sample size. Let's consider an illustration of how a tolerable error and the risk of exceeding it might be set in a given problem.

Example 7.15 New York City receives per capita aid from New York State at the rate of \$6.75 per person. The aid is based on the

population of the city as reported in the most recent census. In 1957 the city felt that its population had increased significantly since 1950. It, therefore, decided to finance a special census in the hope of qualifying for additional state aid.

Actually, a complete count was conducted. However, let us assume that a sample survey had been taken for the same purpose. How large an error could be tolerated? If the tolerable error were as large as 50,000 people, this would mean that the city could be overpaid or underpaid as much as \$337,500. During the 3 years for which the estimate would apply, the over- or underpayment would amount to more than \$1,000,000. It is doubtful whether New York State or New York City would be willing to tolerate an error of this size. A tolerable error more likely would be in the magnitude of at most 10,000 people. The payment would be in error by at most \$67,500 per year. However, neither the city nor the state would like to run the risk of having this tolerable error exceeded very often. They might decide that the risk of exceeding such an error should be .001.

As this illustration indicates, the size of the error that can be tolerated and the risk of exceeding that error that can be taken depend on the nature of the problem and the consequences of an incorrect decision in that problem. Therefore, to set a tolerable error rationally we should determine the consequences of errors of different sizes. Of course, the expected loss resulting from an error in an estimate may depend on subjective values. It may not be possible to measure it explicitly (as could be done in Example 7.15). Thus, an evaluation of consequences may depend to a great extent on the judgment of management. Nevertheless, to insure the correct statement of objectives it is necessary to focus on the consequences of possible errors.

Thus, the necessary sample size depends upon a predetermined tolerable error, which will be denoted by e , and the specified risk of exceeding that error.

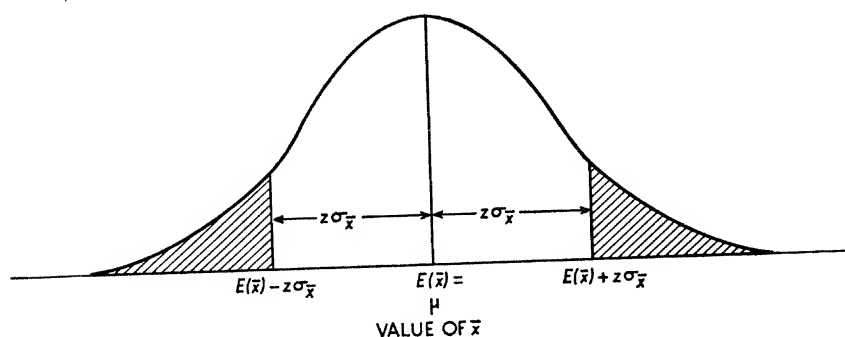
The required sample size also depends upon the estimator selected and the method of sampling to be utilized. For example, suppose that we select a simple random sample and estimate μ by the estimator

$$\hat{\mu} = \bar{x}.$$

We know that this estimator, $\bar{x}$, is approximately normally distributed with its mean equal to $E(\bar{x}) = \mu$ and its standard error equal to

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(assuming simple random sampling from an infinite population). Thus, its sampling distribution appears as follows:



A tolerable error and the risk of exceeding this error may be interpreted easily in terms of this sampling distribution. As we know, $z\sigma_{\bar{x}}$ represents a deviation of a sample mean from μ —it represents an “error” in a value of $\bar{x}$. Hence, the tolerable error in an estimation problem may be equated to $z\sigma_{\bar{x}}$ as follows:

$$e = z\sigma_{\bar{x}}.$$

Since $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ with simple random sampling from an infinite population, we may write

$$e = z \frac{\sigma}{\sqrt{n}}$$

and solve this for n , the sample size, as follows:

$$\sqrt{n} = \frac{z\sigma}{e}$$

$$n = \frac{z^2\sigma^2}{e^2}.$$

This formula enables us to compute the required sample size in a decision problem in which we desire to estimate μ .

What is the value of e ? This is the tolerable error which must be set in each decision problem and which depends on the nature of the problem.

What is the value of z ? First note that the risk of committing an error equal to or greater than e or $z\sigma_{\bar{x}}$ graphically appears as the shaded area in the preceding chart. Of course, this area may be larger or smaller, depending on the value of z —the normal deviate. But the value of z is determined by the risk specified in an estimation problem. For example, if we were to set this risk as .05, z must be 1.96 because 5 per cent of the area under the normal curve

is outside of the interval the mean plus and minus 1.96 standard errors. Other levels of risk which are used often and the corresponding values of z are as follows:

Risk	z
.10.....	1.64
.05.....	1.96
.025.....	2.24
.01.....	2.58
.005.....	2.81

Lastly, to determine the necessary sample size in a problem involving the estimation of μ by the formula $n = z^2\sigma^2/e^2$, we must know σ^2 , the population variance. This often is a problem for it requires a priori information about a population we are planning to study. Nevertheless, there are several methods which can be used to estimate the variance of a population a priori. Two may be noted here. First of all, a similar study may have been conducted recently and an approximation of σ^2 may be derived from such data. Fortunately, the variability in many populations does not tend to change over time. For example, the sales of large retail stores tend to remain large, while those of small stores tend to remain small. Therefore, even though the level of retail sales may change over time, the variability in sales remains approximately the same. Hence, data from last year's study might be used to approximate σ^2 for this year's study. Secondly, estimates of σ^2 may be based on a pilot study—a small study preliminary to the main sample study. In any event, the validity of any assumption about the value of σ^2 and, hence, about the precision a sample yields may be checked after the sample is drawn. In the next section we shall see how this is accomplished.

Thus, once the tolerable error (e) has been set, the risk of exceeding this error determined and translated into a normal deviate (z), and an approximation to σ^2 derived, we may solve for the required sample size.

Example 7.16 In deciding upon a magazine in which to place an advertisement, it is desirable to know the average number of readers per copy. Assume that to obtain this information for a magazine, a statistical survey, using simple random sampling, is to be conducted. Let us assume that the tolerable error in the estimate of the average has been set as .05 persons per copy and that the risk to be taken in having this error exceeded is .01. Further assume that the magazine in question has a circulation of approximately 1,000,000 copies and that the standard deviation of the population is estimated

to be 2 persons. The size of a simple random sample for such a study can be determined as follows:

$$\begin{aligned} n &= \frac{z^2 \sigma^2}{e^2} \\ n &= \frac{(2.58)^2 (2)^2}{(.05)^2} = \frac{26.6256}{.0025} \\ n &= 10,650. \end{aligned}$$

You will note that although N is finite, equal approximately to 1,000,000, we have used the formula to determine n for an infinite population. When N is large, as it is here, this formula yields approximately the correct answer. We shall check our results below by a more precise formula for finite N .

The formula $n = z^2 \sigma^2 / e^2$ applies only for problems involving the estimation of μ by the estimator $\hat{\mu} = \bar{x}$ on the basis of a simple random sample from an infinite population. Similar formulas can be developed for other situations, including other methods of sampling, in much the same way.

If the population being sampled is finite, you can see that we would require a slightly smaller sample than in the infinite case in order to achieve the same precision. To derive a formula for determining n in the case of a finite population, again we begin with

$$e = z\sigma_{\bar{x}},$$

but now we let

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

If we solve

$$e = z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

for n , we find

$$n = \frac{z^2 \sigma^2 N}{(N-1)e^2 + z^2 \sigma^2}.$$

Let's apply this result to the problem presented in Example 7.16 where $N = 1,000,000$, $z = 2.58$, $\sigma = 2$, and $e = .05$. Thus, using the formula for n from a finite population, we find

$$\begin{aligned} n &= \frac{(2.58)^2 (2)^2 (1,000,000)}{(999,999)(.05)^2 + (2.58)^2 (2)^2} \\ &= \frac{26,625,600}{2,526.6231} \\ &= 10,538, \end{aligned}$$

which is approximately equal to the answer (10,650) derived by the simpler formula in Example 7.16.

Let us now consider some of the ways n and the factors determining it, the population standard deviation, the tolerable error, and the level of risk, are related. The formula for n indicates that the smaller the population standard deviation, the smaller the size of the required sample. It also shows that the smaller the risk of exceeding the tolerable error, the larger the sample size. This results from the fact that the smaller the risk, the larger will be the value of the normal deviate, z . Finally, it is easy to see that the larger an error one is willing to tolerate, the smaller the required sample; and that the smaller the tolerable error, the larger the required sample size. Thus, for Example 7.16 we found a sample of 10,538 was required to achieve a tolerable error of .05 persons per copy and a risk level of .01. At this same level of risk, however, a sample of only 2,635 is necessary to achieve a tolerable error of .10 persons per copy—by doubling a tolerable error we may reduce the required sample size to one quarter of its original size. Further, if the tolerable error is cut to one fifth of its original size, the required sample is increased twenty-five fold. These remarks illustrate the general proposition that the sample size required in a study *varies inversely with the square of the tolerable error*.

Careful study should therefore be given to the size of the tolerable error that is set for a study. Although precision is, of course, highly desirable, precision that is greater than required is achieved only at great expense.

You also should note that in setting the risk of exceeding the tolerable error, we implicitly set risks for exceeding other errors as well. Thus, again refer to Example 7.16 and note that a sample of approximately 10,500 is required to keep the risk of exceeding the tolerable error of .05 persons per copy at the .01 level. Of course, the risk of exceeding an error of .04 persons per copy or .03 persons per copy or less is greater than .01, while the risk of exceeding an error of .06 persons per copy or .07 persons per copy or more is less than .01. In general, for a given sample size, the smaller the error, the greater the risk of exceeding that error.

Formulas for determining the required size of a simple random sample in problems involving the estimation of π or A or any other decision-parameter may be derived in a way similar to those for the estimation of μ . (You might try to derive one or two in exactly the same way we derived the formula for the sample size needed to

estimate μ .) Thus, to determine the sample size necessary for the estimation of π we let

$$e = z\sigma_p$$

with

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

and find

$$n = \frac{z^2\pi(1-\pi)}{e^2}.$$

This assumes, of course, simple random sampling from an infinite population. For a finite population the formula for n is

$$n = \frac{z^2N\pi(1-\pi)}{(N-1)e^2 + z^2\pi(1-\pi)}.$$

These formulas for n depend on the assumption that the normal curve adequately describes the sampling distribution under consideration.

To apply either of these formulas for problems involving the estimation of π , we must know π —the very thing we wish to estimate! Again, however, note that a rough idea of the value of π may suffice to determine the necessary sample size. In contrast, the purpose of the study itself is to estimate π more precisely and objectively. Thus, the validity of any assumption about the value of π and, hence, the precision of a result may be checked on the basis of the sample.

Example 7.17 Consider the problem of estimating the proportion of families in New York City who are listening to a given television program. Suppose that the tolerable error is set at 4 percentage points and the risk of having that error exceeded is set at .05. To determine the size of a sample needed to obtain the specified degree of precision, we must assume some approximate value for π . Thus, on the basis of past experience, we might assume π approximately equals .20. N is rather large, and, hence, we may determine the sample size by the formula

$$n = \frac{z^2\pi(1-\pi)}{e^2}.$$

Since the risk is .05, z equals 1.96. Furthermore, e has been set at .04 and π estimated as .20. Thus,

$$\begin{aligned} n &= \frac{(1.96)^2(.20)(.80)}{(.04)^2} \\ &= 384. \end{aligned}$$

Our result indicates that a sample size of about 384—say 400 for convenience—should yield an estimate of π within $\pm .04$ with only a 5 per cent risk of exceeding that error.

It also may be noted that if we lack any other way of approximating the value of π a priori, we may assume that the population is as variable as possible—that $\pi = .5$. This makes $\pi(1 - \pi)$ as large as possible and, hence, for a specified error and risk requires the largest sample size. Hence, this assumption guarantees that the sample size will be large enough to yield a desired tolerable error and risk, although it may make it far larger and more costly than necessary.

The formulas for determining the sample size necessary to estimate decision-parameters require that one center on a tolerable error and a risk of exceeding that error in the planning stage of a statistical study. This is an objective approach to the problem of determining sample size. Of course, after one sets an error and a risk and determines n , it may appear that the cost of the sample is prohibitive, or it may appear that the cost is only moderate and could be increased. Hence, in order to control directly the cost of a study, one common method is to allocate a fixed amount of money for a study and to select as large a sample as possible for that sum. However, it must be remembered that any specified sample size will involve some error and some risk. An arbitrarily set amount of money, then, may mean too much error and risk—it may be better to spend more or give up the study entirely. Alternately, it may mean too little error and risk—less money could be spent. Isn't it better, then, to focus on error and risk first to see what sample size is necessary to achieve these goals? As we know, if the cost of attaining such an error and risk are prohibitive, either the specified error or the risk might be adjusted.

7.5 The Precision Attained

After a sample has been selected, the pertinent observations made, and estimates of the decision-parameters computed, it is possible to estimate the precision or reliability of the estimating procedure from the sample itself. That is, we can estimate, on the basis of the sample, the standard errors of the estimates prepared.

For example, to estimate the average price of a product we would set up a statistical decision rule which would include the sample size necessary to achieve a specified tolerable error and risk of exceeding that error. But the sample size, as we saw in the last section,

is determined on the basis of an assumption about the value of the population standard deviation (or the population proportion). However, this assumption may not be completely accurate. Hence, subsequent to the selection of the sample, an estimate of the standard error of the estimated mean or proportion will indicate the actual precision of our estimating procedure.

The computation of the attained precision of an estimate is important whenever the estimate is to be used by someone other than the persons conducting the study. The standard error of the estimate gives, in this case, an objective measure of the variability of the sampling procedure and, hence, of the reliability of the results. Thus, a jurist in evaluating sample results submitted in evidence may use the standard error to evaluate the procedure by which the results were obtained. As we already know, no one ordinarily can evaluate a single sample result in itself.

The attained precision also is important to those conducting a sample study as a basis for their own decisions. Thus, if the standard error of an estimate turns out too large and, hence, indicates that the estimates are not as precise as they should be, a supplementary sample may be chosen. Estimates then would be based on the combined sample.

As we know, the variance of the sample mean, with simple random sampling, is

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \frac{N - n}{N - 1}.$$

This measure of the precision of a sampling procedure may be estimated from the sample simply by estimating σ^2 . As was noted in Example 7.8, an unbiased estimator of the population variance is

$$\hat{\sigma}^2 = \frac{N - 1}{N} \cdot \frac{n}{n - 1} s^2$$

or, approximately,

$$\hat{\sigma}^2 = \frac{n}{n - 1} s^2,$$

where s^2 is the sample variance. This formula may be used to estimate σ^2 on the basis of a sample and that value used, in turn, to estimate $\sigma_{\bar{x}}^2$ by the formula

$$\hat{\sigma}_{\bar{x}}^2 = \frac{\hat{\sigma}^2}{n} \frac{N - n}{N - 1}.$$

This is an unbiased estimator of $\sigma_{\bar{x}}^2$. Let's apply it.

Example 7.18 Reconsider the sample study planned in Example 7.16 and assume that a simple random sample of 10,538 subscribers to the magazine has been selected in order to estimate the average number of readers per copy. The sample mean is 3.4 readers per copy, which we may take as an estimate of μ .

In addition, we may estimate $\sigma_{\bar{x}}$ as a measure of the precision of the sampling procedure. Thus, suppose that we find the variance of the sample to be 3.6. An estimate of the population variance on the basis of this sample statistic is as follows:

$$\begin{aligned}\sigma^2 &= \frac{n}{n-1} s^2 \\ &= \frac{10,538}{10,537} \times 3.6 \\ &= 3.6.\end{aligned}$$

(As you can see, for a large sample size the estimated population variance is approximately equal to the sample variance.) For an estimate of $\sigma_{\bar{x}}^2$ we thus find (ignoring the f.p.c.)

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \frac{\sigma^2}{n} \\ &= \frac{3.6}{10,538} \\ &= .000342.\end{aligned}$$

Hence, the estimated standard error is

$$\begin{aligned}\sigma_{\bar{x}} &= \sqrt{.000342} \\ &= .018,\end{aligned}$$

which is an estimate of the precision attained in the study.

In Example 7.16 the tolerable error in the estimate of μ was set as .05 readers per copy with a risk of exceeding this error planned as .01. The determination of the sample size necessary to meet these objectives was based on the assumption that $\sigma = 2$. Now, from the sample we have estimated

$$\sigma^2 = 3.6$$

and, hence,

$$\begin{aligned}\sigma &= \sqrt{3.6} \\ &= 1.9.\end{aligned}$$

In view of the fact that this estimate approximately equals the value for σ which we assumed to determine the necessary sample size, it appears that our assumption was reasonable and that our study approximately achieved the planned precision.

For problems in which π is estimated by the sample proportion of a simple random sample, the variance of the estimating procedure is, as we know,

$$\sigma_p^2 = \frac{\pi(1-\pi)}{n} \frac{N-n}{N-1}.$$

This may be estimated from a sample⁴ simply by using p as an estimate of π . Thus,

$$\hat{\sigma}_p^2 = \frac{p(1-p)}{n} \frac{N-n}{N-1}.$$

Example 7.19 In Example 7.17 we found that we needed a simple random sample of about 400 families in order to estimate the proportion of families in New York City who were listening to a given television program with a tolerable error of 4 percentage points and a .05 risk of exceeding that error. Assume that this sample is drawn and that it is found that 30 per cent of the sample families were listening. This 30 per cent provides an estimate of π , the population proportion.

The attained precision of this study can be estimated by using $p = .30$ as an estimate of the population proportion. Thus, an estimate of σ_p is (again ignoring the f.p.c.)

$$\begin{aligned}\hat{\sigma}_p &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{(.30)(.70)}{400}} \\ &= \sqrt{.000525} \\ &= .023.\end{aligned}$$

In contrast, we assumed that $\pi = .2$ in order to determine the required sample size. If we were to measure the precision on the basis of our assumption, it would be

$$\begin{aligned}\sigma_p &= \sqrt{\frac{(.20)(.80)}{400}} \\ &= \sqrt{.000400} \\ &= .020.\end{aligned}$$

Hence, we estimate that the attained precision is slightly less than the planned precision because the value of σ_p estimated from the sample is greater than the value of σ_p that we assumed in determining the required sample size. This further indicates that the actual risk of exceeding the tolerable error of 4 percentage points may be slightly higher than the planned risk of .05.

7.6 Interval Estimation

In the first section of this chapter we distinguish between a point estimate and an interval estimate. The latter, as you will recall, takes the form of a range of estimates about the value of a decision-

⁴ An unbiased estimator of σ_p^2 would ordinarily include $n-1$ rather than n in the denominator. This is a small difference, and the use of n has become common in statistical practice.

parameter. For example, we might estimate that the average age of a group of employees is between 38 and 46 years. This range of values is an interval estimate.

Estimates in the form of intervals usually are called *confidence intervals* in the field of statistics. You see, the purpose of any interval estimate is to try to present a range of values which will include the true value of the decision-parameter under consideration. However, since interval estimates are based on sample information, we cannot expect them to be infallible. Nevertheless, a given procedure for constructing such intervals based on probability samples has a certain probability of including the true value of the decision-parameter. This probability may be taken as a measure of our confidence in the procedure used. Hence, the term "confidence interval" is adopted.

For example, the interval of 38 to 46 years, referred to above, may or may not include the true mean age in the population. However, suppose we knew that if the procedure used for constructing that interval estimate were to be used repeatedly, 90 per cent of the resulting intervals would include the true mean. This relative frequency or probability of .90 would measure our confidence in the procedure used. Hence, it may be referred to as a *confidence coefficient*.

How can we construct interval estimates? How can we measure the confidence that we may have in a given procedure? To answer these questions let us see what would happen in a given situation if we repeated such an estimation procedure over and over again.

Suppose that we were studying the cold cereal consumption of all families in the United States and that we were interested in estimating the average of this population (in terms of ounces per month). Thus, a sample of size 900 is to be selected and an interval estimate of μ computed. The minimum value of the interval is to be computed by the formula

$$\bar{x} - 1.96\sigma_{\bar{x}},$$

while the maximum value of the interval is to be computed by the formula

$$\bar{x} + 1.96\sigma_{\bar{x}}.$$

Let us see what results the above procedure would yield if it were applied repeatedly to a given population. Assume, then, that we know $\mu = 400$ and $\sigma = 120$ for the population. If this is the case,

then the sampling distribution of the sample mean for samples of size 900 would be normally distributed, with $E(\bar{x}) = 400$, and with

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{120}{\sqrt{900}} \\ &= 4\end{aligned}$$

(ignoring the finite population correction factor which should be approximately equal to 1 in this problem).

If a sample of 900 were drawn from such a population and if $\bar{x} = 405$, we would construct the following interval:

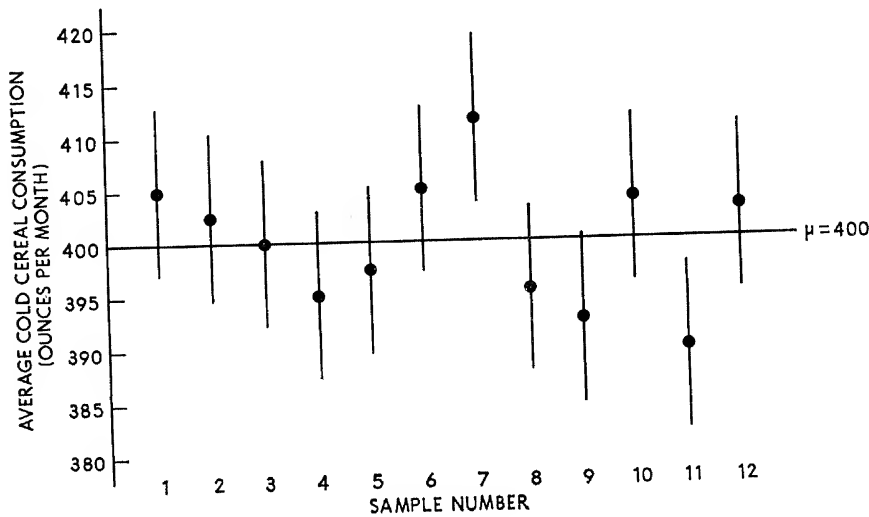
$$\begin{aligned}\bar{x} - 1.96\sigma_{\bar{x}} &= 405 - (1.96)(4) & \bar{x} + 1.96\sigma_{\bar{x}} &= 405 + (1.96)(4) \\ &= 405 - 7.84 & &= 405 + 7.84 \\ &= 397.16, & &= 412.84.\end{aligned}$$

This interval of 397.16 to 412.84 includes the population mean ($\mu = 400$) of our assumed population. However, if our sample mean had been 410, the interval would have been

$$\begin{aligned}\bar{x} - 1.96\sigma_{\bar{x}} &= 410 - (1.96)(4) & \bar{x} + 1.96\sigma_{\bar{x}} &= 410 + (1.96)(4) \\ &= 402.16, & &= 417.84.\end{aligned}$$

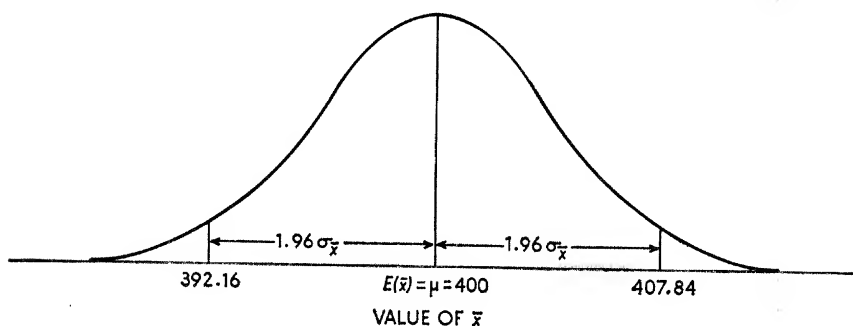
This interval would not include the true population mean.

The diagram below illustrates a series of possible intervals. Each dot represents a sample mean; each vertical line represents the range of the resulting interval estimate.



The horizontal line (at $\mu = 400$) indicates the population mean. Thus, of the 12 intervals diagrammed, we can see that only 2 fail to include the population mean; that is, only 2 do not intersect the horizontal line at $\mu = 400$.

More fully, you can see from the following graph that any sample mean that is at a distance of $1.96\sigma_{\bar{x}}$ (7.84 ounces) or less from the population mean, in either direction, will lead to intervals that include the population mean. That is, only values of $\bar{x}$ between 392.16 and 407.84 will result in an estimate that includes the popu-



lation mean ($\mu = 400$). Any sample mean less than 392.16, or more than 407.84, will lead to intervals that do not include the population mean.

Now, suppose that we were to set up interval estimates by repeatedly drawing samples of 900 from our assumed population (with $\mu = 400$ and $\sigma = 120$). We can reason that 95 per cent of these intervals would include the population mean. This is so because 95 per cent of the sample means will be included between the population mean plus and minus $1.96\sigma_{\bar{x}}$ (between 392.16 and 407.84). Only 5 per cent of the sample means will fall outside of this interval. Hence, only 5 per cent will lead to interval estimates that do not include $\mu = 400$.

The 95 per cent confidence interval is a consequence of our use of $1.96\sigma_{\bar{x}}$ in computing an estimate. The 1.96, you will recognize, is a normal deviate. In a similar manner, by choosing the proper normal deviate, we can set up intervals with any degree of confidence that we desire. Thus, the use of a normal deviate equal to 2.58 involves a risk of .01—and, hence, a confidence interval of .99. A .99 confidence interval is one derived by a procedure which if repeated indefinitely would result in only .01 of our interval estimates not including the population mean.

In an actual statistical problem of estimation, we do not know the population mean. Therefore, we have no way of knowing whether the one interval, $\bar{x} \pm 1.96\sigma_{\bar{x}}$, that we set up includes or does not include the population mean. However, we do know that the interval which we set up was determined by a procedure which if repeated again and again would yield a set of intervals, .95 of which would contain the population mean.

Another problem also presents itself. We ordinarily do not know σ , and, therefore, we cannot compute $\sigma_{\bar{x}}$. However, we may estimate σ and hence $\sigma_{\bar{x}}$ on the basis of the sample in the way considered in the last section. Thus, an interval estimate for μ is given by

$$\bar{x} - z\hat{\sigma}_{\bar{x}} \quad \text{and} \quad \bar{x} + z\hat{\sigma}_{\bar{x}}.$$

The normal deviate (z) is governed by the choice of the confidence coefficient.

Example 7.20 A simple random sample of 100 is drawn by a trade association from its membership of 500 firms in order to estimate the average hourly wage for milling machine operators. A .95 confidence interval is to be calculated. It is found that the sample mean is \$2.48 and the standard deviation of the population is estimated at \$.50. The confidence interval is given by

$$\bar{x} - z\hat{\sigma}_{\bar{x}} \quad \bar{x} + z\hat{\sigma}_{\bar{x}}$$

with $z = 1.96$ because a .95 confidence interval is to be computed. The standard error of the mean may be estimated as follows:

$$\begin{aligned}\hat{\sigma}_{\bar{x}} &= \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \\ &= \frac{.50}{\sqrt{100}} \sqrt{\frac{400}{499}} \\ &= .045.\end{aligned}$$

Thus, the confidence interval is

$$\begin{array}{rcl} 2.48 - (1.96)(.045) & 2.48 + (1.96)(.045) \\ 2.48 - .09 & 2.48 + .09 \\ 2.39 & 2.57 \end{array}$$

We thus estimate that the population mean is included in the interval, \$2.39–\$2.57. Of course, we do not know whether this one particular interval contains the population mean, μ . However, we can measure the reliability of the procedure used in computing this interval estimate. Its reliability is indicated by the confidence coefficient of .95. If many intervals, such as the \$2.39 to \$2.57, were computed by repeating our sampling procedure over and over again, about 95 per cent of such intervals would include the true value of μ .

The theory of confidence intervals for other decision-parameters is quite similar to that sketched here. Thus, for an interval estimate of π we simply compute

$$p - z\hat{\sigma}_p \quad \text{and} \quad p + z\hat{\sigma}_p,$$

which give the lower and upper limits of the interval.

Example 7.21 Suppose that the simple random sample of 100 firms considered in the previous illustration also is used to estimate the proportion, π , of firms working 3 shifts. Forty of the 100 sample firms operated 3 shifts at the time of the survey, so that $p = .40$. Suppose that we compute a .99 confidence interval estimate of π . This confidence interval is given by

$$p - z\hat{\sigma}_p \qquad p + z\hat{\sigma}_p$$

with $z = 2.58$. The formula for estimating σ_p , the standard error of a sample proportion, is

$$\begin{aligned}\hat{\sigma}_p &= \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}} \\ &= \sqrt{\frac{(.40)(.60)}{100}} \sqrt{\frac{400}{499}} \\ &= .044.\end{aligned}$$

The confidence interval is

$$\begin{array}{ll}.40 - (2.58)(.044) & .40 + (2.58)(.044) \\ .40 - .11 & .40 + .11 \\ .29 & .51\end{array}$$

Thus, we estimate that the population proportion is between .29 and .51. This is one of many such intervals that could be derived by the same sampling procedure—approximately 99 per cent of such intervals would include the true population proportion.

7.7 Estimation of Distributions

In some statistical decision problems, as was pointed out in Chapter 3, a decision maker would like to know the whole frequency distribution of a population. In this section, we shall consider the use of the normal curve as a population model in connection with the estimation of population frequency distributions. Thus, we shall assume that the population of interest may be adequately described by the normal distribution. The normal curve then may be used to find the necessary frequencies by measuring areas under the normal curve.

The estimation of a distribution by means of the normal curve is

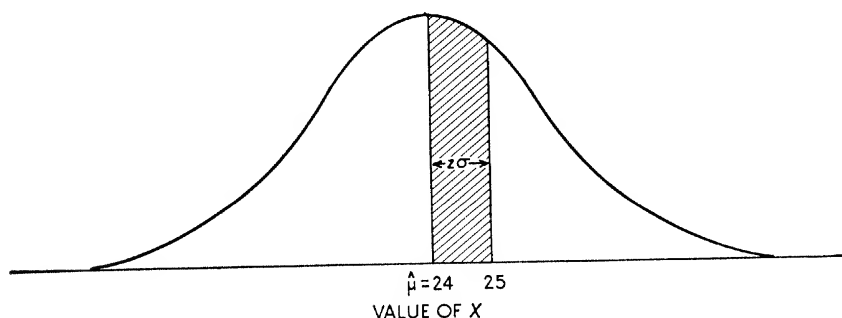
a relatively simple procedure. A normal curve, as we know, is defined by only two parameters—its arithmetic mean (μ) and its standard deviation (σ). One need not, in such a case, determine all of the many possible irregularities in the exact population frequency distribution. To derive estimates of μ and σ is sufficient.

Example 7.22 A company which manufactures automobile batteries considers several types of guarantees that it might offer to customers. One suggestion is to offer a complete refund on any batteries that fail within 1 year coupled with a partial refund on any that fail during the 13th to the 24th month. Another suggestion includes a double-your-money-back offer. All involve some set of offers scaled to the month in which the battery fails.

The decision problem is to pick one guarantee. Officials of the company, however, are uncertain about the expected cost of each of their alternatives. In order to calculate the expected cost, per battery, for each guarantee they must know the number of batteries which are expected to fail in the 23rd month from date of purchase, in the 24th month from date of purchase, etc. That is, the firm must know the frequency distribution of the times-to-failure of the batteries its manufacturing process can turn out.

Suppose it can be assumed that this population of times-to-failure is normally distributed. In order to calculate an approximation to the frequency distribution based upon this assumption, only the mean and standard deviation of the population need be determined. Thus, suppose it were estimated that $\mu = 24$ and $\sigma = 2$ months.

The relative frequency of failures expected between 24 and 25 months (the 25th month of use) can be determined by measuring the following area under the normal curve:



In order to determine this area we compute the normal deviate

$$\begin{aligned} z &= \frac{X - \mu}{\sigma} \\ &= \frac{25 - 24}{2} \\ &= 0.5. \end{aligned}$$

The table of areas under the normal curve indicates a tail area (past 25 months) of .3085 for $z = 0.5$. The central area (24 to 25 months) is thus

$$.5000 - .3085 = .1915,$$

which represents the relative frequency of failures expected between 24 and 25 months.

Similarly, we may determine that the relative frequency of failures expected between 24 and 26 months (the 25th and 26th months of use) is .3413 (the area under the normal curve from μ to $\mu + \sigma$). Thus, the relative frequency of those failing in the 26th month (between 25 and 26 months) is

$$\begin{array}{r} .3413 \text{ (failing between 24 and 26 months)} \\ - .1915 \text{ (failing between 24 and 25 months)} \\ \hline .1498 \end{array}$$

Furthermore, to find the relative frequency of those expected to fail in the 20th month (between 19 and 20 months), they would note: (a) .4938 is the area within the interval 19 to 24 months ($\mu - 2.5\sigma$ to μ); (b) .4772 is the area within the interval 20 to 24 months ($\mu - 2\sigma$ to μ); and, hence (c), the area between 19 and 20 months is $.4938 - .4772 = .0166$.

Similar calculations would lead the firm to the following relative frequency distribution (which you should verify):

Expected Month of Failure	Months of Use	Relative Frequency
17th.....	16 and under 17	.0002
18th.....	17 " " 18	.0011
19th.....	18 " " 19	.0049
20th.....	19 " " 20	.0166
21st.....	20 " " 21	.0440
22nd.....	21 " " 22	.0919
23rd.....	22 " " 23	.1498
24th.....	23 " " 24	.1915
25th.....	24 " " 25	.1915
26th.....	25 " " 26	.1498
27th.....	26 " " 27	.0919
28th.....	27 " " 28	.0440
29th.....	28 " " 29	.0166
30th.....	29 " " 30	.0049
31st.....	30 " " 31	.0011
32nd.....	31 " " 32	.0002
		1.0000

In applying the normal population model, or any other model, in the way outlined, it must be remembered that unless the model used is a good replica of the population, the estimated frequency distribution it provides will be considerably biased. Thus, if one is

thought necessary, a model must be selected carefully on the basis of the combined judgments and experiences of management and the statistician. There also exist certain statistical methods for determining the adequacy of models, but they are beyond the scope of this text.

7.8 You Should Now Know That

A statistical estimation problem involves selecting, on the basis of sample information, an estimate which approximates the value of a decision-parameter.

A point estimate takes the form of a single value; an interval estimate the form of a range of values.

A decision rule in an estimation problem must specify a method of sample selection, the size of the sample, and an estimator.

An estimator is a formula which translates sample observations into an estimate of a decision-parameter.

Among the statistical criteria for evaluating estimators are consistency, unbiasedness, and efficiency.

The choice of a method of sampling depends on both the cost and the precision of available methods.

One method of determining sample size in an estimation problem is to specify a tolerable error in an estimate and the risk of exceeding that error, and then solve for the required n by the appropriate formula.

After the sample has been drawn and an estimate calculated, it is desirable also to estimate the standard error of that estimate. This is an indicator of the attained precision.

A confidence interval is a range of values which, it is hoped, includes the true value of a decision-parameter. Any one interval estimate may or may not include the true value. However, the confidence coefficient, e.g., 95 per cent or 99 per cent, measures the probability that the procedure used will result in an interval actually including the true value.

When it is appropriate, the normal population model may be used in the estimation of a population frequency distribution. This requires only the estimation of μ and σ and the computation of frequencies by measuring areas under the normal curve.

7.9 You Should Now Be Able to Solve

1. A statistics instructor conducts a sampling experiment in class. A simple random sample of 5 students is drawn from the class

of 40. The average weight of the students in the sample is 160.2 pounds. This is an estimate of the population (class) average. The instructor also requests one student to select a judgment sample of 5 students. The result is a sample mean of 153.8 pounds which is another estimate of the population (class) average.

Subsequently, a complete survey of the class shows the true value of the population mean to be 155.1 pounds. Should the instructor and the class conclude that the mean of a judgment sample is a better estimator than the mean of a simple random sample for estimating the average weights of students in statistics classes? Why or why not?

2. Which method of sample selection (simple random, stratified, or cluster) would you recommend in the following situations? Explain your answers.

- a) A study to indicate the attitude of factory workers toward the establishment of a company cafeteria.
- b) A study to estimate the number of vacant dwelling units in a city whose population is over 1 million.
- c) A study to estimate the proportion of defective toothbrush handles produced by a given manufacturing process.
- d) A study to estimate the inventory of a department store.
- e) A study to estimate the proportion of high school seniors in a given state who plan to enter college.

3. You are to conduct an opinion poll to determine the opinions of residents of a given community about a projected industrial development program. How large a sample should you select to estimate the proportion of adult residents favoring the projected development? Make all assumptions necessary to determine the sample size, and justify these assumptions.

4. An aircraft parts manufacturer wishes to determine the average shearing strength of a certain type of weld in order to submit a bid for a contract to produce these parts. What size sample is required so that the risk of exceeding an error of 20 pounds or more is .005? Assume that σ is 100 pounds.

5. The market research department of an electric appliance manufacturing concern has been allocated \$500 to conduct a survey in a given state to determine the proportion of retailers who prefer a new design for a kitchen clock to the current design. The firm is to use this information as a guide in deciding whether or not to market the clock. For the estimate to be of use it is desirable that the error should not exceed 5 percentage points.

For the amount budgeted, can the market research department provide the firm with an estimate of the required precision? Explain your answer. (Assume that the cost of an interview is \$1.25, that π is about .50, and that simple random sampling is to be employed.)

If you were the department head, what suggestions would you make to management?

6. A ratio-delay study is conducted to determine the proportion of time a key-punch operator spends in unproductive work. A ratio-delay study is conducted by making observations at instants of time that are selected at random. At each instance the key-punch operator will be or will not be engaged in productive activity.

- a) How large a sample is needed to determine the proportion of unproductive time so that an error of 3 percentage points is not exceeded more than 1 time in a hundred? (Assume π is .10.)
- b) How large a sample would be needed so that the risk of exceeding an error of 6 percentage points or more is .01?
- c) If we assume π is .20, would a larger or smaller sample be required? Explain.
- d) If we are willing to risk an error of 3 percentage points or more 5 times in a hundred, will the sample size be smaller or larger than in (a)? Explain.

7. A sample of 500 accounts receivable is selected from the 4,032 accounts that a firm has, and the sample mean is found to be \$242.30. The sample standard deviation is computed to be \$3.20. Set up a .99 confidence interval estimate of the population mean. How do you interpret the meaning of this interval?

8. A survey on consumer finances reports that 33 per cent of a sample of 2,600 spending units expected good times during the next 12 months. Assume that a simple random sample was used in the study. Set up a .95 confidence interval estimate of the population proportion of spending units expecting good times.

7.10 You Will Also Find That

For further discussions of the topics covered in this chapter and for some supplementary ideas on estimation, the following are key references:

- HIRSCH, WERNER Z. *Introduction to Modern Statistics*, pp. 125-37, 154-58. New York: Macmillan Co., 1957.
- ROSANDER, A. C. *Elementary Principles of Statistics*, pp. 229-309. New York: D. Van Nostrand & Co., Inc., 1951.
-

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics: A New Approach*, pp. 443-70. Glencoe, Ill.: The Free Press, 1956.

NETER, JOHN, and WASSERMAN, WILLIAM. *Fundamental Statistics for Business and Economics*, pp. 307-47. New York: Allyn & Bacon, Inc., 1956.

Risk and Uncertainty in Testing Problems

8.1 Risks in Testing Problems

In the preceding chapter we considered the problem of formulating decision rules for *estimation* problems. In this chapter we shall develop methods for formulating decision rules for *testing* problems. A testing problem, as you should remember, is one in which a decision maker must choose between two or more qualitative courses of action. A testing problem can be treated statistically if, and only if, the choice of an alternative course of action depends exclusively on knowledge of the value of some decision-parameter(s). For simplicity in our introductory survey of statistics we shall concentrate on *two-action* testing problems—that is, on those problems in which the decision maker has only two alternatives to choose from. For example, accepting or rejecting a lot of raw materials, marketing or not marketing a new product, and hiring or not hiring an applicant for an executive training program are all illustrations of two-action testing problems. Remember, however, that multi-action testing problems often arise in the business world.

One basic question to be answered in this section is: “What do we mean by risks in a testing problem?” We have already answered this question here and there throughout the first seven chapters. But because this material is here and there and not centralized in any one section, we had better co-ordinate it right now. To review certain ideas about risks in testing problems, let us consider the case of an advertising manager who must decide whether or not to place a par-

ticular advertisement in a local newspaper. Suppose that he were to flip a coin to reach his decision. If a head appeared, he would place the advertisement, and if a tail appeared, he would not. Then, if an advertisement were a potentially effective one (and, therefore, should be inserted), the advertising man would be running a 50 per cent risk of making the wrong decision (not inserting the advertisement). Furthermore, if the advertisement were a potentially ineffective one, the coin-flipping decision procedure would result in a 50 per cent risk of making the wrong decision (inserting an ineffective advertisement).

In short, this procedure and the probabilities associated with it might be summarized as follows:

Alternative	State of the World	
	Effective	Ineffective
Place the ad.	Correct decision .50	Incorrect decision .50
Do not place the ad. . .	Incorrect decision .50	Correct decision .50

Notice that all the probabilities are conditional probabilities—they depend upon the state of the world. For example, the first .50 tells us that *if* the advertisement is a potentially effective one, the coin-flipping procedure would give us a .50 chance of making the correct decision—placing the advertisement. The other probabilities are to be interpreted similarly.

Of course, flipping a coin to make advertising decisions is a naïve procedure. The risks entailed are great. A company in the long run would have a small chance of survival if half of the potentially ineffective advertisements were placed and half of the potentially effective advertisements were not used. In fact, no advertising director is needed for such a decision procedure. His training and experience are not drawn upon.

By contrast, an advertising manager, before making a decision, might desire to select a sample to determine the proportion of people who see and remember the advertisement. This procedure reduces his uncertainty and the risks of making errors in the classification of advertisements. But to go to another extreme, note that every advertising director would like to be able to boast about the following probabilities:

Alternative	State of the World	
	Effective	Ineffective
Place the ad.	Correct decision 1	Incorrect decision 0
Do not place the ad. . .	Incorrect decision 0	Correct decision 1

This situation means that a correct decision would always be made. While this is a highly desirable situation, it is ordinarily impossible to achieve when decisions are made in the face of uncertainty. The only way in which the incorrect decision of neglecting to place potentially effective advertisements can be avoided is to place every advertisement. But, this would mean that we could never avoid the error of placing a potentially ineffective advertisement. Similarly, the only way we could avoid the placing of an ineffective advertisement would be not to place any advertisement. This, however, would never give an effective advertisement the possibility of being inserted.

Although, it is impossible to eliminate either type of incorrect decision, statistical theory enables us to control the relative frequencies of these errors. The level at which these risks are to be controlled will depend upon the consequences of an incorrect decision and the costs of sampling. Thus, the consequences of placing an ineffective advertisement involve the needless expenditure of funds to insert the advertisement and perhaps a loss of sales. The consequences of neglecting to place an effective advertisement involve the expense of preparing the advertisement and the potential loss of sales because the advertisement was not inserted. The risks decided upon should be those that result in minimum economic losses when the costs of sampling and the consequences of both types of errors are taken into consideration.

After such an analysis of risks and costs, the advertising director might decide that he is willing to assume the following risks:

Alternative	State of the World	
	Effective	Ineffective
Place the ad.	Correct decision .95	Incorrect decision .10
Do not place the ad. . .	Incorrect decision .05	Correct decision .90

Thus, the advertising manager would be satisfied with a decision procedure for which the probability of neglecting to place an effective advertisement is .05 and the probability of placing an ineffective advertisement is .10.

These two types of incorrect decisions have special statistical names. They are called *errors of the first kind* and *errors of the second kind*. They also are referred to as *Type I errors* and *Type II errors*, respectively. By convention, we designate as the error of the first kind (Type I) the error that we are most eager to avoid. When we are just as eager to avoid both types of error, generally either error may be referred to as the Type I error. In addition, the probability of incurring a Type I error is termed the α -risk (*alpha risk*), and the probability of incurring a Type II error, the β -risk (*beta risk*). Thus, the error of neglecting to place an effective advertisement is a Type I error since this is the risk the advertising manager is more eager to avoid and the probability of the occurrence of such an error, the α -risk, is .05. The error of placing an ineffective advertisement is the Type II error, and the probability of its occurrence—the β -risk is .10. Notice that a Type I error can occur only *if* the advertisement is effective and a Type II error can occur only *if* the advertisement is ineffective. Therefore, the probability that an error will occur is always conditional upon the existence of a particular state of the world.

The advertisement classification problem as we have formulated it cannot be treated statistically. To do so, it is necessary to describe the states of the world in terms of a decision-parameter. Suppose that the advertising manager selects as the decision-parameter the proportion of readers who see and remember the advertisement. The states of the world, therefore, could be described by values of π ranging from .00 (none remember) to 1.00 (all remember). However, the statistical technique that we shall employ to develop decision procedures requires that we focus our attention on only two of the many possible descriptions of the state of the world—i.e., two values of π . Thus, the advertising manager might select $\pi = .25$ as the point at which he would like to be almost sure of using the advertisement. In other words, $\pi = .25$ describes a state of the world which is associated with a potentially effective advertisement. To describe a state of the world associated with a potentially ineffective advertisement the manager might select $\pi = .10$ —a point at which he would hesitate to place the advertisement. We can restate the problem as follows:

Alternative	State of the World	
	$\pi = .25$	$\pi = .10$
Place the ad.	Correct decision .95	Incorrect decision .10
Do not place the ad. . .	Incorrect decision .05	Correct decision .90

The advertising manager thus indicates that he would like to insert an advertisement which 25 per cent of the readers will see and remember. However, after considering the costs of sampling and the costs of incorrect action, he is willing to run the risk of not placing such advertisements only 5 times in 100. On the other hand, he would prefer not to place advertisements if only 10 per cent of the readers will remember them. But after his consideration of the costs of sampling and of incorrect action, he is willing to run the risk of placing such advertisements 10 times in 100.

The aspects of testing problems that we have been considering—the specification of alternative courses of action, the selection of a decision-parameter, the particular values of the decision-parameter upon which to focus, and the specification of the α - and β -risks—are primarily responsibilities of management. However, the statistician usually must take an active part in their selection. He must provide estimates of sampling costs and assist in the determination of the consequences of both types of error. It then becomes the job of the statistician to formulate a decision rule, the application of which will control risks of incorrect action at the specified probabilities. In the next section we consider the development of such decision procedures.

8.2 Formulating the Decision Rule—the Proportion as a Decision-Parameter

We have seen that one way in which management sets its requirements for a decision rule is to demand that the rule offer a requisite degree of protection against making incorrect decisions. In statistical terms, management is looking for a decision rule whose application will offer assurance that: (a) the probability of a Type I error will be α , and (b) the probability of a Type II error will be β . In this section we shall consider a method for determining such a rule for problems in which the population proportion is the decision-

parameter; that is, for problems in which π adequately describes states of the world. Let's consider the following example to explain the method.

Example 8.1 A marketing decision must be made on whether or not to market a new scouring powder. A sample of consumers is to be taken and the decision between marketing the product or not marketing it is to be based on this sample.

Management and the statistician decide to concentrate on the proportion of family units who say that they prefer this scouring powder after using a free sample for a week—this proportion for all family units is the decision-parameter. They further focus attention on two values of π : $\pi_0 = .25$, at which point it would be preferable *not* to market the new product; and $\pi_1 = .40$, at which point it would be preferable to market.

In summary, the decision problem is as follows:

Alternative	State of the World	
	$\pi_0 = .25$	$\pi_1 = .40$
Market.....	Incorrect decision	Correct decision
Don't market.....	Correct decision	Incorrect decision

One other preliminary consideration involves a statement of the risks that management is willing to take. Suppose that they are willing to take only a .10 chance (the β -risk) of not marketing the product if $\pi = .40$. They also are willing to take only a .05 chance (the α -risk) of marketing the product if $\pi = .25$.

The Type I error in this problem is the marketing of the scouring powder *if* $\pi = .25$ since this is the error management is more eager to avoid. The probability of incurring this error is .05 as against the probability of .10 of incurring the Type II error—not marketing the product *if* $\pi = .40$.

Before considering the determination of the decision rule we shall introduce some statistical terminology that often is used in discussions of testing problems. This terminology, while not essential to an explanation of the concepts involved in testing problems, is widely used in the field.

The two values of π upon which we center our attention represent assumptions that we make concerning the state of the world. Statisticians call such assumptions *hypotheses*. In the problem we are considering, one hypothesis is that $\pi = .40$; the other is that $\pi = .25$. If we market the new scouring powder, the hypothesis $\pi = .40$ is accepted and the hypothesis $\pi = .25$ is rejected. On the

other hand, if we do not market the scouring powder, the hypothesis $\pi = .40$ is rejected and the hypothesis $\pi = .25$ is accepted. Thus, it is possible to describe the action taken in terms of only one of the hypotheses since the acceptance of one means the rejection of the other, and vice versa. The hypothesis upon which we center our attention is called the *null hypothesis* and is designated as H_0 ; the other hypothesis is known as the *alternate hypothesis* and is designated as H_1 .

It is customary to designate as the null hypothesis that hypothesis which, if rejected when true, leads to a Type I error. Thus, in the problem under consideration, $\pi_0 = .25$ is the null hypothesis. Why? If $\pi = .25$ and the scouring powder is marketed, a Type I error will result. Obviously, $\pi_1 = .40$ is the alternate hypothesis. The probability of a Type I error (α) is thus the probability of rejecting the null hypothesis if it is true. The β -risk represents the probability of accepting the null hypothesis if the alternate hypothesis is true. Problems of the type that we are considering, therefore, can be described as tests of hypotheses—to accept the null hypothesis is to take one action, and to reject the null hypothesis is to take the other.

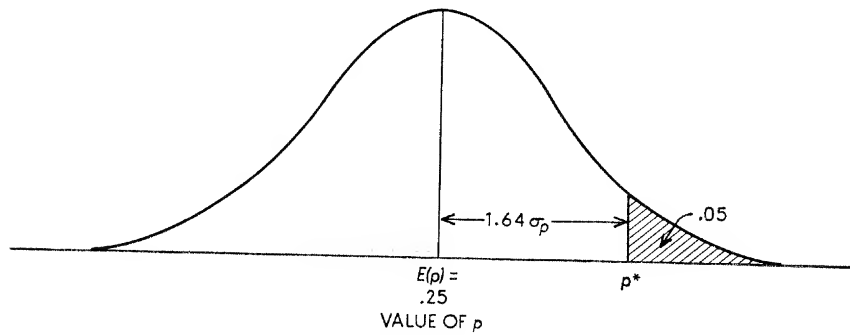
We return now to the formulation of a decision rule for the marketing problem under consideration. Generally, the rule will take the following form: Select a simple random sample of size n and observe the sample proportion, p . If p is less than or equal to some critical value (which we shall denote by p^*), do not market the product; if p is greater than this critical value, market the product. Observe carefully the essential elements of a decision rule for a testing problem:

1. A method of sample selection (e.g., simple random sampling),
2. The sample size (n), and
3. The critical value of the statistic (e.g., p^*) which governs the action to be taken for various values of the statistic.

Since simple random sampling is assumed, our task is to find values of n and p^* for the decision problem under consideration.

This decision rule must offer a degree of protection against the occurrence of two types of error. We may now consider these errors one at a time. A Type I error can occur only if $\pi = .25$ (if the null hypothesis is true). However, we know that the normal curve will approximate the sampling distribution of p , and that if $\pi = .25$, its mean, $E(p)$, also is .25. In this sampling distribution we need to

find a critical value, p^* , so that .05 of the area of the normal curve is to the right of p^* as in the diagram below:



Why .05? If $\pi = .25$, management is willing to assume only this degree of risk in marketing the product. Since the product will be marketed when we observe a value of p greater than p^* , only 5 per cent of all possible sample proportions should exceed p^* if $\pi = .25$.

The table of areas under the normal curve indicates that we must go out $1.64\sigma_p$ to reach a point in the tail beyond which .05 of the area is contained. This tells us what value p^* must equal if our predetermined risk of .05 is to be met. Thus

$$p^* = .25 + 1.64 \sigma_p.$$

But,

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}.$$

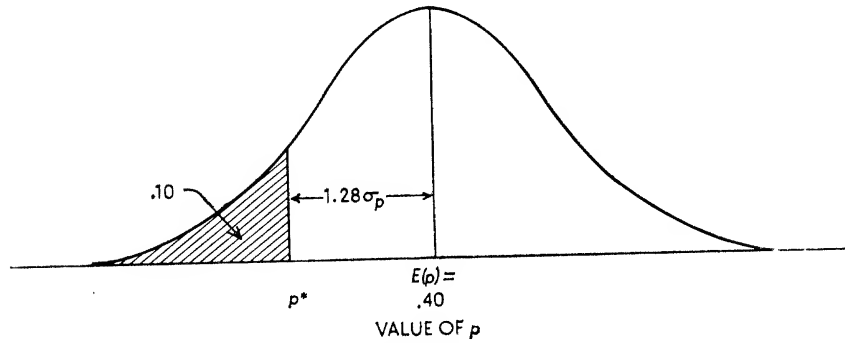
Hence,

$$\begin{aligned} p^* &= .25 + 1.64 \sqrt{\frac{\pi(1 - \pi)}{n}} \\ &= .25 + 1.64 \sqrt{\frac{(.25)(.75)}{n}} \\ &= .25 + \frac{.710}{\sqrt{n}}. \end{aligned}$$

This is the relationship that p^* and n must satisfy to meet the condition that the probability of a Type I error is .05.

We now consider the Type II error. This error can occur only if $\pi = .40$ (if H_1 , the alternate hypothesis is true). Again we know that the sampling distribution of p will be approximated by the normal curve, and if $\pi = .40$, $E(p)$ also is .40. For this distribution

we want to find another equation relating p^* and n so that .10 of the area is to the left of p^* , as in the diagram below:



Why .10? If $\pi = .40$, management is willing to take a risk of .10 of not marketing the product. The product will not be marketed if p is equal to or less than p^* . Therefore, only 10 per cent of the possible sample proportions should have values equal to or less than p^* if $\pi = .40$.

The table of areas under the normal curve indicates that if we go out $1.28\sigma_p$ from $E(p)$, the area in the tail is .10. This means that

$$\begin{aligned} p^* &= .40 - 1.28\sigma_p \\ &= .40 - 1.28\sqrt{\frac{\pi(1-\pi)}{n}} \\ &= .40 - 1.28\sqrt{\frac{.4 \times .6}{n}} \\ &= .40 - \frac{.626}{\sqrt{n}} \end{aligned}$$

This is the relationship which p^* and n must satisfy to meet the condition that the probability of a Type II error is .10.

Earlier we also found that

$$p^* = .25 + \frac{.710}{\sqrt{n}}$$

The values of p^* and n which will satisfy both of these equations will be the basis for our decision rule. We can solve for n by equating both expressions for p^* as follows:

$$\begin{aligned} .40 - \frac{.626}{\sqrt{n}} &= .25 + \frac{.710}{\sqrt{n}} \\ .15 &= \frac{.710 + .626}{\sqrt{n}} \\ \sqrt{n} &= \frac{1.336}{.15} \\ \sqrt{n} &= 8.91 \\ n &= 79.3, \text{ or } 80. \end{aligned}$$

Thus, the required sample size is 80. We can determine the value of p^* by substituting the value of n in either of the two expressions for p^* .

$$\begin{aligned} p^* &= .40 - \frac{.626}{\sqrt{n}} & p^* &= .25 + \frac{.710}{\sqrt{n}} \\ &= .40 - \frac{.626}{8.91} & &= .25 + \frac{.710}{8.91} \\ &= .40 - .07 & &= .25 + .08 \\ &= .33 & &= .33 \end{aligned}$$

We now have determined the decision rule. It is: Select a simple random sample of 80 families and determine the proportion of families in the sample that prefer the scouring powder. If 33 per cent or less favor the scouring powder, do not market it. If more than 33 per cent prefer the scouring powder, market it. This rule meets the risks that management and the statistician agreed upon.

In general, the procedure we have followed is: We focus attention on two states of the world, π_0 (the null hypothesis) and π_1 (the alternate hypothesis). We determine the risks we are willing to take in the decision problem. The general equations for determining the values of n and p^* are then

$$\begin{aligned} p^* &= \pi_0 + z_\alpha \sqrt{\frac{\pi_0(1 - \pi_0)}{n}} \\ p^* &= \pi_1 - z_\beta \sqrt{\frac{\pi_1(1 - \pi_1)}{n}}. \end{aligned}$$

z_α and z_β are the appropriate number of standard deviations—the appropriate normal deviates—required to yield the predetermined α - and β -risks. These two equations are solved to determine the sample size (n) and the critical value of p (p^*).

8.3 Formulating a Decision Rule—the Mean as a Decision-Parameter

The previous section dealt with situations in which the population proportion is used to describe states of the world. We now shall consider the formulation of a decision rule for a problem in which the population mean is the decision-parameter. We shall relate our discussion to the following problem situation.

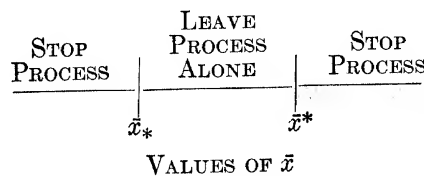
Example 8.2 A firm packages instant coffee in small-sized jars—net weight 2 ounces. It is interested in controlling its production process; in making sure that each of the jars contains about 2 ounces.

The company has studied its process when it was operating satisfactorily. It has estimated that the average contents per jar is 2.06

ounces, and the standard deviation of the process is .02 ounces. The company is satisfied with this performance and wants to maintain it in future production.

We shall assume throughout our discussion that σ remains at .02 ounces, and develop a decision rule to control variations in μ . (To control variations in σ we would need a second decision rule.) Thus, assume that the company plans to take a sample of n jars every hour. On the basis of the average weight of these n jars it will decide whether to leave the process alone or to stop the process in order to try to correct whatever is wrong.

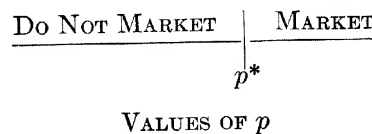
In addition to the use of the sample mean as a basis for decision, this situation differs in another important way from the one we considered in the preceding section. In this problem the process is operating unsatisfactorily if the jars are either overfilled or underfilled. Therefore, a decision rule must divide all possible values of the sample mean into three regions, as illustrated in the following diagram:



where $\bar{x}_*$ represents the lower critical value of the sample mean and $\bar{x}^*$ the upper critical value of the sample mean.

The diagram indicates that very high values (presumably overfilling) and very low values (presumably underfilling) of the statistic lead to the same action, while middle values lead to the alternate action. Since both extremes—high and low values—lead to the same course of action, we have what is called a *two-tailed* testing problem.

In contrast, let us reconsider the decision rule developed to decide whether or not to market a new scouring powder. That rule divides possible values of the statistic, p , into two regions: values equal to or less than p^* , and values greater than p^* . Any value within one of these regions leads to a unique action, as shown in the diagram below:



Since only the higher or the lower values are involved in a decision, we have what is called a *one-tailed* testing problem.

When should a decision rule be one-tailed and when should a de-

cision rule be two-tailed? The answer is relatively simple. We ask ourselves: "Suppose that we knew the decision-parameter in the problem under consideration; would we take action A for high *and* low values of this parameter or would we take that action only for high *or* low values?" If the answer is high *and* low, we have a two-tailed testing problem, otherwise we have a one-tailed testing problem.

To return to the coffee jar problem, let us consider the general form of the statistical decision rule we must formulate. This form will be as follows: Select a simple random sample of n jars and determine the average weight of their contents. If this average weight is between $\bar{x}_*$ and $\bar{x}^*$, let the process continue to run. If the average is equal to or less than $\bar{x}_*$, or if it is equal to or greater than $\bar{x}^*$, stop the process.

To determine the values of n , $\bar{x}_*$, and $\bar{x}^*$ for this rule it is necessary for management first to focus attention upon values of the population mean to describe states of the world. The company, as noted earlier, wishes to leave the process alone if $\mu = 2.06$, the normal process average, and selects this as one value of μ on which to focus attention. The company would want to stop the process if the value of μ departed from 2.06 by an amount which would result in excessive overfilling or underfilling. It thus chooses $\mu = 2.03$ and $\mu = 2.09$ as the values of the decision-parameter which describe the alternative states of the world. The 2.03 represents a state in which the process is underfilling and, hence, a state for which the company would like to stop the process. The 2.09 represents a state in which the process is overfilling and, hence, also a condition which the company would like to try to remedy by stopping and checking the process. Notice that $\mu = 2.03$ and $\mu = 2.09$ are both .03 ounces away from the value of $\mu = 2.06$ which describes the first state of the world. This symmetrical formulation simplifies the determination of a decision rule.

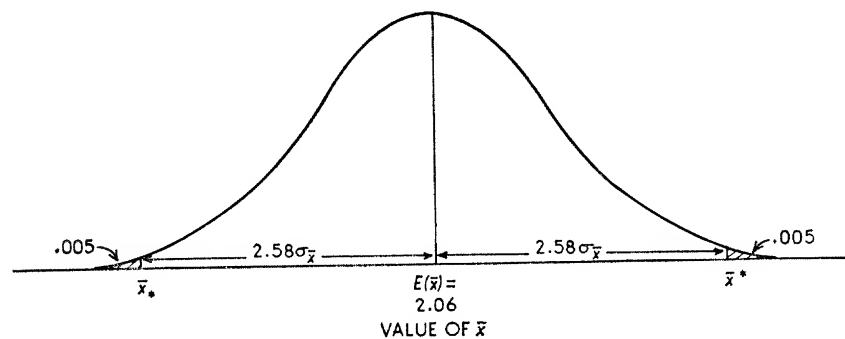
Our decision problem can be summarized, and risks assigned, as follows:

Alternative	State of the World		
	$\mu_0 = 2.06$	$\mu_1 = 2.03$	$\mu_2 = 2.09$
Leave process alone	Correct decision .99	Incorrect decision .05	Incorrect decision .05
Stop process	Incorrect decision .01	Correct decision .95	Correct decision .95

This indicates that the company is willing to take only a .01 risk of stopping the process if $\mu = 2.06$ (if H_0 is true). Since the company is more eager to avoid the error of stopping the process if $\mu = 2.06$, this may be regarded as the Type I error. Hence, the .01 is the α -risk. Our summary further indicates that the company is willing to take only a .05 risk of leaving the process alone if $\mu = 2.03$; that is, if there is excessive underfilling. It also is willing to take a .05 risk of leaving the process alone if $\mu = 2.09$; that is, if there is excessive overfilling. Note that the alternate hypothesis, H_1 , refers to the possibility that μ equals either 2.03 or 2.09. Hence, there are two β -risks—one if $\mu = 2.03$, the other if $\mu = 2.09$.

Let us now determine the values of n , $\bar{x}_*$, and $\bar{x}^*$ for the required decision rule. This rule must assure that the risks of incorrect decision are as specified above. We shall consider these risks one at a time.

We first consider the α -risk, the probability of leaving the process alone if $\mu = 2.06$. This has been set as .01. Note that the sampling distribution of the sample mean is approximated by the normal curve, and if $\mu = 2.06$, its mean, $E(\bar{x})$, also is 2.06. For this sampling distribution, we want to find the critical values, $\bar{x}_*$ and $\bar{x}^*$, so that .01 of the area of the normal curve falls beyond these values as in the diagram below:



The area has been divided equally between the two tails because of the symmetry of the problem. If $\mu = 2.06$, the error that may occur is to stop the process. The process will be stopped only if the value of a sample mean does *not* fall between $\bar{x}_*$ and $\bar{x}^*$. Since only .005 of all possible sample means will be equal to or more than $\bar{x}^*$, and only .005 will be equal to or less than $\bar{x}_*$, the probability of stopping the process, if $\mu = 2.06$, will be .01—the risk set by the company.

The table of areas under the normal curve indicates that since

the area in the tail is .005, the distance from $E(\bar{x})$ to either of the critical $\bar{x}$ values is $2.58\sigma_{\bar{x}}$. Thus,

$$\bar{x}_* = 2.06 - 2.58\sigma_{\bar{x}} \quad \text{and} \quad \bar{x}^* = 2.06 + 2.58\sigma_{\bar{x}}.$$

But,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Hence,

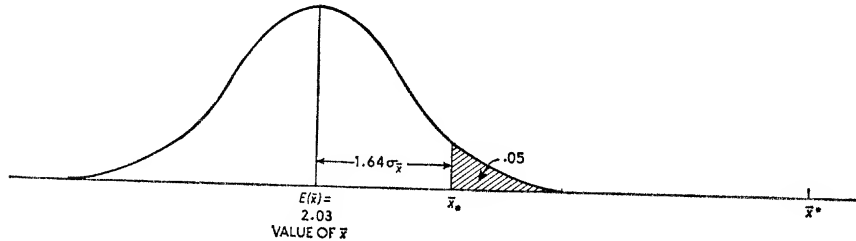
$$\bar{x}_* = 2.06 - 2.58 \frac{\sigma}{\sqrt{n}} \quad \bar{x}^* = 2.06 + 2.58 \frac{\sigma}{\sqrt{n}}.$$

Furthermore, as pointed out in Example 8.2, σ has been estimated as .02. Therefore,

$$\begin{aligned} \bar{x}_* &= 2.06 - (2.58) \frac{(.02)}{\sqrt{n}} & \bar{x}^* &= 2.06 + (2.58) \frac{(.02)}{\sqrt{n}} \\ &= 2.06 - \frac{.0516}{\sqrt{n}} & &= 2.06 + \frac{.0516}{\sqrt{n}}. \end{aligned}$$

These are the relationships which n , $\bar{x}_*$, and $\bar{x}^*$ must satisfy to meet the conditions that the probability of stopping the process if $\mu = 2.06$, is .01.

We now direct our attention to the second risk—the probability of leaving the process alone if $\mu = 2.03$ (excessive underfilling). If $\mu = 2.03$, the sampling distribution of the mean will be approximated by the normal curve with $E(\bar{x}) = 2.03$. In this sampling distribution only .05 of the possible sample means must be greater than $\bar{x}_*$, as illustrated in the diagram below:



Why .05? The company is willing to assume only this risk of leaving the process alone if $\mu = 2.03$. Since the process will be left alone when we observe a value of $\bar{x}$ between $\bar{x}_*$ and $\bar{x}^*$, only 5 per cent of all possible means should exceed $\bar{x}_*$. (If $\mu = 2.03$, practically no sample means will be greater than $\bar{x}^*$. Therefore, we can say that the proportion of possible sample means between $\bar{x}_*$ and $\bar{x}^*$ is the same as the proportion greater than $\bar{x}_*$.)

Since the area beyond $\bar{x}_*$ is .05, it is $1.64\sigma_{\bar{x}}$ from $E(\bar{x})$. Therefore,

$$\bar{x}_* = 2.03 + 1.64\sigma_{\bar{x}}.$$

But,

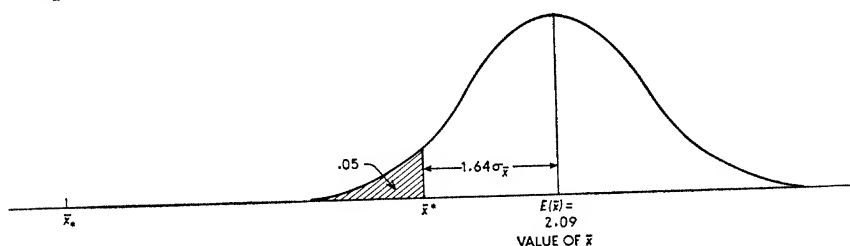
$$\sigma_{\bar{x}} = \frac{.02}{\sqrt{n}}.$$

Hence,

$$\bar{x}_* = 2.03 + \frac{.0328}{\sqrt{n}}.$$

This is the relationship n and $\bar{x}_*$ must satisfy to fulfill the condition that the probability of leaving the process alone, if $\mu = 2.03$, is .05.

We now consider the error of leaving the process alone if $\mu = 2.09$ —i.e., if there is excessive overfilling. The sampling distribution of the mean is again approximated by the normal curve, and $E(\bar{x}) = 2.09$. In this sampling distribution, only .05 of the possible sample means must be less than $\bar{x}^*$, as shown in the diagram below:



The risk of leaving the process alone if $\mu = 2.09$ is set at .05—only .05 of all possible sample means should be less than $\bar{x}^*$. (If $\mu = 2.09$, practically no sample means will be less than $\bar{x}_*$. Therefore, we can say that the proportion of all possible sample means that fall between $\bar{x}^*$ and $\bar{x}_*$ is the same as the proportion less than $\bar{x}^*$.)

Since the area beyond $\bar{x}^*$ is .05, it is $1.64\sigma_{\bar{x}}$ from $E(\bar{x})$. Thus,

$$\bar{x}^* = 2.09 - 1.64\sigma_{\bar{x}}.$$

But,

$$\sigma_{\bar{x}} = \frac{.02}{\sqrt{n}}.$$

Hence,

$$\bar{x}^* = 2.09 - \frac{.0328}{\sqrt{n}}.$$

This is the relationship that n and $\bar{x}^*$ must satisfy to meet the condition that the probability of leaving the process alone, if $\mu = 2.09$, is .05.

We now have two expressions for each of the critical values, $\bar{x}_*$ and $\bar{x}^*$, namely,

$\bar{x}_*$ Equations	$\bar{x}^*$ Equations
$\bar{x}_* = 2.06 - \frac{.0516}{\sqrt{n}}$	$\bar{x}^* = 2.06 + \frac{.0516}{\sqrt{n}}$
$\bar{x}_* = 2.03 + \frac{.0328}{\sqrt{n}}$	$\bar{x}^* = 2.09 - \frac{.0328}{\sqrt{n}}$

We can equate each pair to solve for n .

$2.06 - \frac{.0516}{\sqrt{n}} = 2.03 + \frac{.0328}{\sqrt{n}}$	$2.09 - \frac{.0328}{\sqrt{n}} = 2.06 + \frac{.0516}{\sqrt{n}}$
$.03 = \frac{.0844}{\sqrt{n}}$	$.03 = \frac{.0844}{\sqrt{n}}$
$n = 7.9, \text{ or } 8$	$n = 7.9, \text{ or } 8$

We can now substitute in either of the equations for $\bar{x}_*$ and $\bar{x}^*$ to solve for the critical values.

$\bar{x}_* = 2.06 - \frac{.0516}{\sqrt{n}}$	$\bar{x}^* = 2.06 + \frac{.0516}{\sqrt{n}}$
$= 2.06 - \frac{.0516}{\sqrt{8}}$	$= 2.06 + \frac{.0516}{\sqrt{8}}$
$= 2.06 - .02$	$= 2.06 + .02$
$= 2.04$	$= 2.08$

You will observe that both sets of equations yield the same value for n and that both critical values deviate by .02 ounces from 2.06 (μ_0). It is sufficient, therefore, to consider equations for either $\bar{x}_*$ or $\bar{x}^*$ to determine the decision rule. This simplified result follows from the symmetrical ways in which we selected states of the world and set the risks (μ_1 and μ_2 were .03 ounces below and above μ_0 , and both β -risks were set at .05). This symmetrical formulation of the problem simplifies the determination of a decision rule and is sufficient for most practical purposes.

In summary, the statistical decision rule for the problem formulated in Example 8.2 is as follows: Select a simple random sample of 8 jars and determine the average weight of their contents. If this average weight is between 2.04 and 2.08, let the process continue to run. If the average is equal to or less than 2.04, or if it is equal to or greater than 2.08, stop the process.

In general, a procedure used to determine a decision rule for a two-tailed testing problem is as follows: Focus attention on three states of the world, μ_0 (the null hypothesis), and μ_1 and μ_2 (the alternate hypothesis). Let μ_1 and μ_2 be represented by two values equidistant, but in opposite directions, from μ_0 . Set the α -risk and the β -risk. Determine n , $\bar{x}_*$, and $\bar{x}^*$ from the following equations:

$$\begin{aligned}\bar{x}_* &= \mu_0 - z_{\frac{\alpha}{2}} \sigma_{\bar{x}} \\ \bar{x}_* &= \mu_1 + z_{\beta} \sigma_{\bar{x}} \\ \text{and} \\ \bar{x}^* &= \mu_0 + z_{\frac{\alpha}{2}} \sigma_{\bar{x}} \\ \bar{x}^* &= \mu_2 - z_{\beta} \sigma_{\bar{x}}\end{aligned}$$

Remember that $z_{\frac{\alpha}{2}}$ is the value of the normal deviate for an area in the tail equal to one half of α , and z_{β} is the value of the normal deviate for an area in the tail equal to β .

The same procedure and set of equations can be adapted for a two-tailed testing problem in which π is the decision-parameter by substituting π for μ , p_* for $\bar{x}_*$, p^* for $\bar{x}^*$, and σ_p for $\sigma_{\bar{x}}$.

8.4 Operating Characteristic Curves

To develop a decision rule we had to concentrate on two possible values of the decision-parameter. Thus, we centered our attention on $\pi_0 = .25$ and $\pi_1 = .40$ in the marketing decision problem. We centered our attention on $\mu_0 = 2.06$, $\mu_1 = 2.03$, and $\mu_2 = 2.09$ in the coffee jar problem. We decided upon the risks of incorrect action we were willing to take on the assumption that these values of the decision-parameter described the true state of the world. There are, however, many other possible values of the decision-parameter that might be descriptive of the true state of the world. Thus, 10 per cent, or 20 per cent, or 50 per cent of the families might prefer the new scouring powder. There would be risks associated with taking action on the basis of the derived decision rule if any of these values of the decision-parameter were to describe the true state of the world.

Thus, we want a method of indicating the probability of taking a certain action for a given decision rule for all possible values of a decision-parameter. A statistical way of presenting these probabilities usually is called an *operating characteristic curve* or, more

briefly, an *OC* curve. It also has been called a performance characteristic curve.

Let us now determine the *OC* curve for the decision rule derived in connection with the marketing of a new scouring powder. To determine an *OC* curve it is customary to compute the probabilities of taking the action which may lead to a Type II error (the probabilities of accepting the null hypothesis). We, therefore, shall consider the probability of *not* marketing the new product under the decision rule for all possible values of π . These values can range from .00 to 1.00. We now recall the provisions of the decision rule: If in a sample of 80 family units the value of p is equal to .33 or less, do not market the scouring powder; otherwise market it.

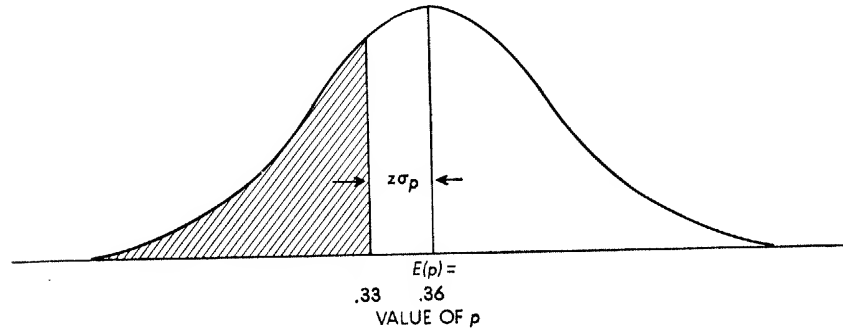
We can determine two points on the *OC* curve very easily. If $\pi = .00$, the probability of *not* marketing the product is 1.00 since every value of p will be equal to .00—less than .33. If $\pi = 1.00$, we will always market the product and, hence, the probability of not marketing it is .00. Two other points follow from the α - and β -risks set in deriving the decision rule. The risk of marketing the product if $\pi = .25$ (α -risk) was set at .05. Therefore, the probability of not marketing the product if $\pi = .25$ is .95. The probability of not marketing the product if $\pi = .40$ (β -risk) was set at .10.

A fifth point that can be determined easily is the probability of *not* marketing if π is equal to .33—the same as the critical value, p^* . The sampling distribution of p if $\pi = .33$ is normally distributed with its mean equal to .33. Since the normal curve is distributed symmetrically about its mean, half of the possible values of p are equal to .33 or less. Hence, under the decision rule the probability of not marketing the scouring powder if $\pi = .33$ is .50. We summarize below the five points of the *OC* curve just derived:

π	Probability of Not Marketing
.00.....	1.00
.25.....	.95
.33.....	.50
.40.....	.10
1.00.....	.00

We now shall illustrate how to determine the probability of not marketing the product for other values of π ; for example, $\pi = .36$. The statistical problem is to determine, given $\pi = .36$, the probability of getting a value of p equal to .33 or less. We have dealt with problems of this type in Chapter 6, when we demonstrated

how to measure risks by means of the normal curve. However, let's review the procedure. The probability in question is illustrated in the diagram below:



To determine the shaded area we compute the normal deviate,

$$z = \frac{p - \pi}{\sigma_p}.$$

But,

$$\begin{aligned}\sigma_p &= \sqrt{\frac{\pi(1 - \pi)}{n}} \\ &= \sqrt{\frac{(.36)(.64)}{80}} \\ &= .054.\end{aligned}$$

Hence,

$$\begin{aligned}z &= \frac{.33 - .36}{.054} \\ &= -.56.\end{aligned}$$

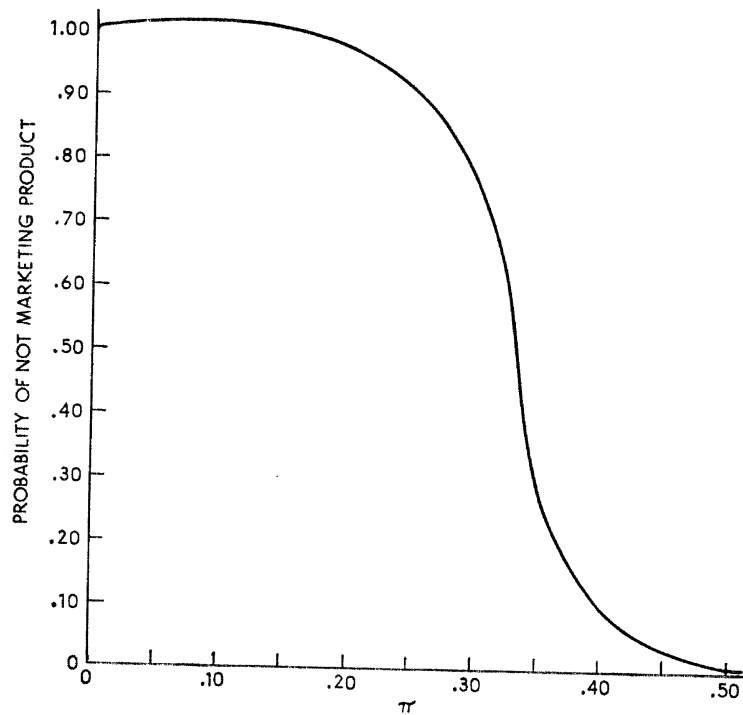
When $z = .56$, the table of areas under the normal curve indicates that the area in the tail is .2877. Hence, the probability of not marketing the scouring powder, if $\pi = .36$, is .2877. Similarly, we could compute the probabilities of not marketing the product for other possible values of π .

Why is an *OC* curve called a curve? Because the type of information we have been computing is usually presented graphically. Thus, for the decision rule we have been considering, the *OC* curve is as shown in Figure 8.1. You will observe that the *OC* curve has only one tail. This is a property of the *OC* curve for any decision rule relating to a one-tailed test.

A little interpretation of the *OC* curve is in order. It shows con-

ditional probabilities. Thus, for example, it indicates that if we use the decision rule under consideration, and if $\pi = .36$, the probability of not marketing the new scouring powder is approximately .29 (or that the probability of marketing it is .71). Similar statements can be made on the basis of the curve for any value of π . The *OC* curve thus gives us a device for evaluating the protection offered by a given decision rule against taking incorrect actions. Note that the

Figure 8.1



larger the value of π , the smaller is the probability of *not* marketing the product, and, hence, the greater is the probability of marketing it. We could not have confidence in a decision procedure which did not exhibit this characteristic.

Let us now develop an *OC* curve for a decision rule in a two-tailed testing problem in which the mean is the decision-parameter. Such a rule was derived in Section 8.3: If, for a simple random sample of 8 jars of coffee, the average weight is between 2.04 and 2.08 ounces, leave the process alone; otherwise, stop the process. For an *OC* curve it is customary to depict the probabilities of taking the action that may lead to a Type II error. Thus, in this case

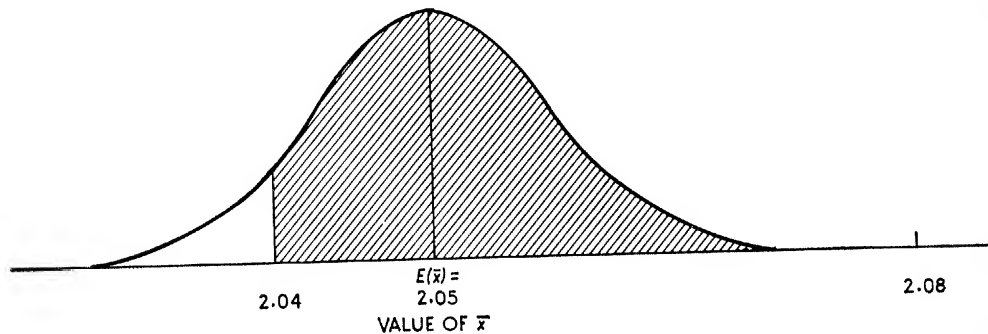
we shall determine the probability of leaving the process alone under various assumptions about μ .

To illustrate the computation of these probabilities suppose that we assume $\mu = 2.05$. The sampling distribution of the mean will be approximated by the normal curve with $E(\bar{x})$ equal to 2.05. The standard error of the mean is

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{.02}{\sqrt{8}} \\ &= .0071.\end{aligned}$$

In developing the decision rule it was stated that the company had estimated σ to be .02 ounces. It was assumed that this value does not vary even though μ does.

The probability of leaving the process alone is represented by the shaded areas in the diagram below:

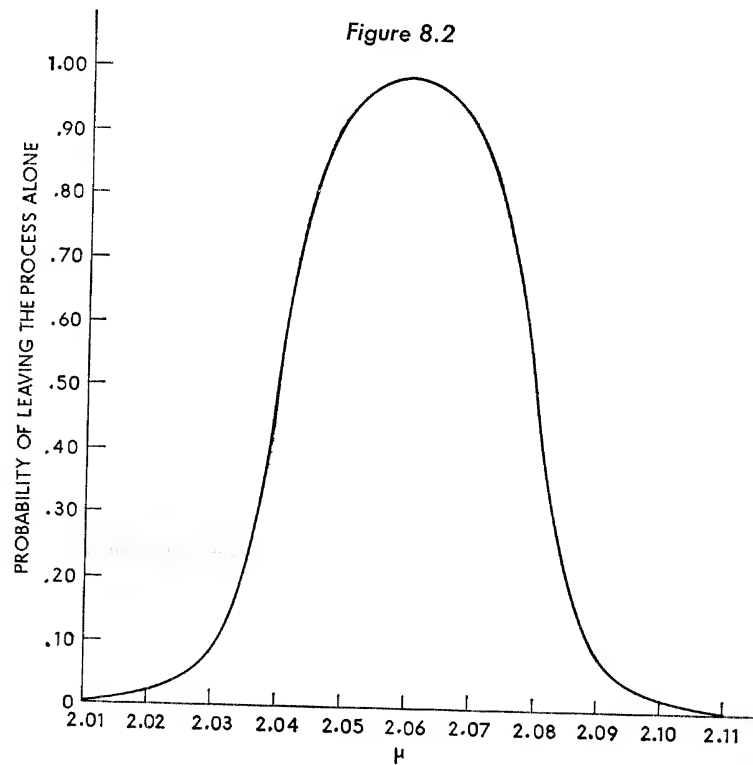


Note that the process is left alone if $\bar{x}$ is between 2.04 and 2.08. To measure the probability of leaving the process alone, we must add the areas in both central portions. To measure areas we must determine the normal deviates. The normal deviates are:

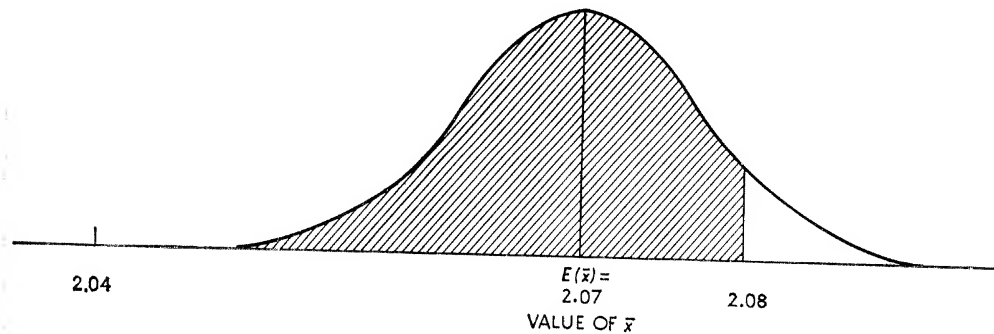
<i>Left Tail</i>	<i>Right Tail</i>
$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$	$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$
$= \frac{2.04 - 2.05}{.0071}$	$= \frac{2.08 - 2.05}{.0071}$
$= -1.41$	$= 4.23$

When z is 1.41, the area in the tail is .0793. Therefore, the area in the left central portion is .4207 (.5000 - .0793). When $z = 4.23$, the area in the tail is practically zero and, hence, the area in the

right central portion may be taken as .50. Therefore, the probability of leaving the process alone if $\mu = 2.05$ is .9207.



Because of the symmetry of the critical values of $\bar{x}$ about 2.06, we would obtain the same probability for leaving the process alone if $\mu = 2.07$. This would reverse the areas in the right and left tails as shown in the diagram below:



In a similar manner the probability of leaving the process alone, given other values of μ , has been computed. You should verify these

probabilities. Again observe the symmetry of these probabilities about $\mu = 2.06$.

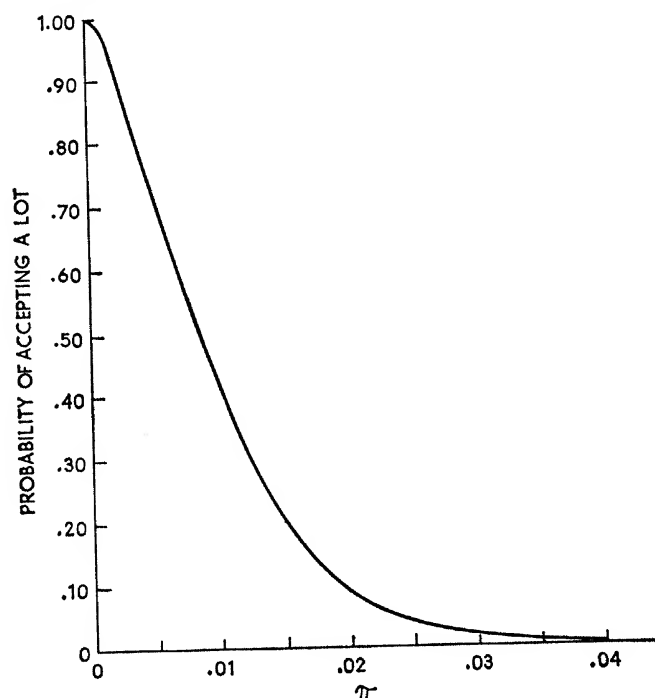
μ	$P(\text{Leaving the Process Alone})$
2.02.....	.0024
2.03.....	.0793
2.04.....	.5000
2.05.....	.9207
2.06.....	.9900
2.07.....	.9207
2.08.....	.5000
2.09.....	.0793
2.10.....	.0024

Figure 8.2 (p. 253) depicts this two-tailed *OC* curve.

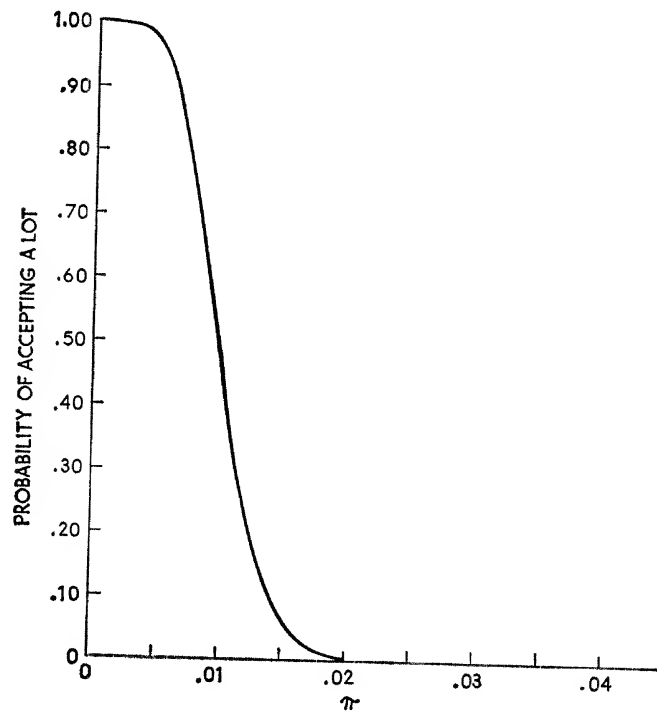
We turn now to an important consideration with respect to *OC* curves—the effect of a change in sample size upon such curves. To illustrate this effect consider the following situation.

Example 8.3 A company accepts or rejects lots of 10,000 machine parts on the basis of the following decision rule: Select a simple random sample of 200 parts. If 1 per cent or more (2 parts or more) are defective, reject the lot; otherwise, accept it.

The operating characteristic curve for this decision rule is shown in the graph below. Notice that π , the proportion of defectives in a lot, is the decision-parameter.

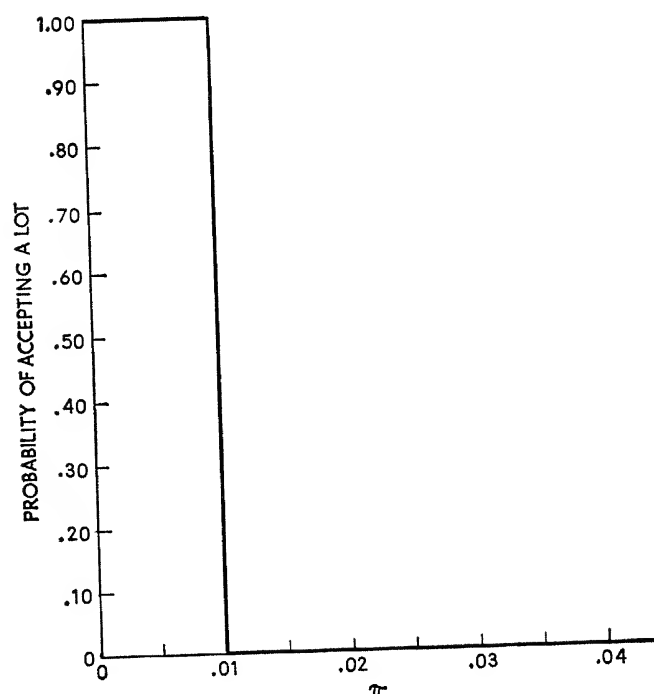


In contrast, suppose that the company modified its decision rule to require a simple random sample of 1,000. Its rule then would be: Select a sample of 1,000 parts. If 1 per cent or more (10 parts or more) are defective, reject the lot; otherwise, accept it. The *OC* curve for the modified rule would appear as:



Notice that this second curve is steeper than the first. This indicates that if the value of π is low, the probability of accepting a lot is larger with the second rule than with the first—that is, the risk of rejecting a good lot is smaller. Furthermore, if the value of π is high, the probability of accepting a lot is smaller with the second rule than with the first—that is, the risk of accepting a poor lot is smaller. In short, the *OC* curve shows that both risks decrease with the second rule. This also is common sense. The second rule involves a larger sample. Hence, it should involve smaller risks.

Thus, as our illustration indicates, the larger the sample, the smaller are all risks. And this reflects itself in a steeper *OC* curve. In fact, suppose that we took a complete count of all 10,000 machine parts in a given lot and rejected the lot only if 1 per cent or more were defective. The *OC* curve for this decision rule would appear as drawn on the following page.



This *OC* curve, based on a complete count, may be considered a limiting case. It represents a decision procedure characterized by certainty: all lots with π equal to .01 or less have a probability of one of being accepted; all others have a probability of zero of being accepted. This limiting case of the *OC* curve can occur only if complete information is obtained. If sample information is used as a basis for decision, there always is some risk of acting incorrectly. The *OC* curve objectively depicts these risks for all possible values of a decision-parameter.

In conclusion, it may be noted that the *OC* curve often is necessary to evaluate the results of secondary studies. For example, suppose that some researcher in the field of education recommends, on the basis of a sample study, that method A of teaching basic statistics be preferred to method B. Among the information about the study that we should seek is the decision rule which led the researcher to his conclusion. Since his conclusion was based on sample information, there are risks that his recommendation was wrong. Hence, the *OC* curve which depicts these risks is an important consideration in our decision on whether or not to follow the recommendation. This choice depends on the risks involved and our attitude towards risk taking.

8.5 Acceptance Sampling

The theory that we have developed in this chapter has been applied most effectively to decisions on managerial problems in the area of inspection. Inspection refers to the function of checking the quality of some work already done. Problems of inspection arise in different areas of the business world. Thus, in manufacturing we have: (1) the inspection of purchased materials, parts, and supplies; (2) the inspection of work at various stages of the manufacturing process; and (3) the inspection of the finished product. Checking the quality of incoming plastic material or inspecting a lot of finished electrical sockets represent such inspection situations.

However, the word "inspection" is also used to refer to the checking or verification of clerical work and records. Thus, reviewing an inventory count or checking payroll vouchers illustrate clerical inspection problems.

The statistical techniques designed to cope with problems of inspection are known as methods of *acceptance sampling*. Acceptance sampling procedures are designed to take action on aggregations of similar product—called a *lot*. The lot can be a case of milk bottles, a bale of cotton, a shipment of wire, the entries in an accounts payable ledger, or a group of punch cards.

The alternative courses of action in acceptance sampling situations are described as (a) accept the lot, and (b) reject the lot. In industrial or clerical problems the action, to accept a lot, usually means that no inspection is needed. The action, to reject a lot, in industrial problems, may mean: (1) send the lot back to the supplier, (2) completely inspect the lot, (3) rework the lot, (4) scrap the lot, or (5) sell the lot as second grade. The nature of the problem in each case dictates the particular type of action to be taken. In clerical operations, rejection usually means completely checking or verifying all items in a lot.

To apply acceptance sampling, it is customary to define two states of the world—good lots and poor lots. We can portray an acceptance sampling decision problem as follows:

Alternative	State of the World	
	Good Lot	Poor Lot
Accept the lot.....	Correct	Incorrect
Reject the lot.....	Incorrect	Correct

The errors, therefore, are to reject a good lot (Type I error) or to accept a poor lot (Type II error). The errors have been so designated to achieve uniformity. In addition, the probability of rejecting a good lot—the α -risk—is known as the *producer's risk*. The probability of accepting a poor lot—the β -risk—is known as the *consumer's risk*.

Generally, the producer's risk is set at a lower value than the consumer's risk. In other words, the decision maker is less eager to reject good product than to accept poor product. This is a result of analyzing the consequences of these risks. In clerical operations, it means that a department head is less eager to accuse an employee unjustly of doing poor work than he is to let a poor worker go by unnoticed. The reasons for this should be obvious. In industry it means that it is less desirable to delay a production schedule and to incur the wrath of a supplier than to rework poor product.

As in any testing problem, we must describe states of the world in terms of a decision-parameter. The decision-parameter used most often in acceptance sampling is the population proportion, π —though μ , σ , and other decision-parameters are used occasionally. We must focus our attention upon two values of the decision-parameter—one value descriptive of good quality and the second value descriptive of poor quality. These values often are given special names. Thus, when π is the decision-parameter, one value (π_0) is called the *acceptable quality level* and is referred to as the *AQL*. It represents that proportion of defectives in a good lot—a lot which a decision maker is willing to reject with a probability of α . The second value (π_1) is the *lot tolerance per cent defective* which usually is referred to as the *LTPD*. It represents the proportion of defectives in a poor lot—a lot which a decision maker is willing to accept with a probability of β .

Example 8.4 A television manufacturer receives shipments of cut wire in lots of about 4,000 pieces. For each lot, he must decide whether to accept it or to reject it. A piece of wire (about 10 inches long) must be of a certain resistance to be usable. Therefore, a wire is either acceptable (if it can be used) or defective (if it cannot be used).

Suppose the company sets the following objectives in its acceptance sampling plan:

$$\begin{array}{ll} \text{AQL: } \pi_0 = 2\% & \alpha = .05 \\ \text{LTPD: } \pi_1 = 6\% & \beta = .10 \end{array}$$

This means that it is willing to use a sampling plan which will achieve the following:

1. If a shipment of wires is 2 per cent defective, the probability it will be accepted is .95 and the probability it will be rejected (α) is .05. This 2 per cent ($\pi_0 = .02$) is the acceptable quality level.
2. If a shipment of wires is 6 per cent defective, the probability it will be rejected is .90 and the probability it will be accepted (β) is .10. This 6 per cent ($\pi_1 = .06$) is the lot tolerance per cent defective.

Table 8.1
TABLE FOR DETERMINING
ACCEPTANCE SAMPLING DECISION RULES
ASSUMING SIMPLE
RANDOM SAMPLING
 $\alpha = .05$ $\beta = .10$

R	c	$n\pi_0$
45.10	0	.051
10.96	1	.355
6.50	2	.818
4.89	3	1.366
4.06	4	1.970
3.55	5	2.613
3.21	6	3.285
2.96	7	3.981
2.77	8	4.695
2.62	9	5.425
2.50	10	6.169
2.40	11	6.924
2.31	12	7.690
2.24	13	8.464
2.18	14	9.246
2.12	15	10.040

Source: F. E. Grubbs, "On Designing Single Sampling Inspection Plans," *Annals of Mathematical Statistics*, Vol. XX, (1949), p. 256.

INSTRUCTIONS FOR USE

1. Calculate $R = \pi_1/\pi_0$ (the ratio of the *LTPD* to the *AQL*).
2. Find R in the table above. If it is not there, use the next smaller value shown.
3. Read off c . This is the acceptance number.
4. Read off $n\pi_0$ and divide by the specified π_0 (the *AQL* for the problem) to get n . This is the required sample size.
5. The statistical decision rule is: Select a simple random sample of size n , and accept the lot only if there are c or fewer defectives in this sample.

Since we have α , β , π_0 , and π_1 , we could determine the decision rule for the above illustration by the methods developed in Section 2 of this chapter. If we were to do so, we would find that n should be about 200 and p^* about .04. Thus, the decision rule is as follows: Select a simple random sample of 200 wires. If $p = .04$ or more, reject the lot; if p is less than .04, accept the lot.

In acceptance sampling instead of expressing the critical value in terms of p we usually convert p^* into an *acceptance number* (c). An acceptance number is the maximum number of defectives we can observe and yet accept the lot. In the decision rule stated above, p^* is .04. In a sample of 200 this represents 8 defective units. Since we would reject a lot if a sample of 200 units had 4 per cent or 8 defectives, the acceptance number is 7—we would still accept the lot if c were 7. This decision rule can be very simply written as ($n = 200$, $c = 7$). This is a concise way of saying: Select a simple random sample of 200 items. If 7 items or less are defective, accept the lot; if 8 items or more are defective, reject the lot. The use of the acceptance number obviates the necessity of converting the number of defectives into a proportion in order to apply a decision rule.

It would be an arduous task if a firm had to use the method of Section 8.2 to determine a decision rule for each of its many inspection operations. Therefore, various general tables have been prepared to facilitate the determination of decision rules. One such table and instructions for its use are presented in Table 8.1. This table is based on a producer's risk (α) of .05 and a consumer's risk (β) of .10. Other tables are available for different values of α and β .

We now illustrate the use of Table 8.1 (p. 259) to formulate a decision rule.

Example 8.5 The work of key-punch operators is to be checked in batches of 2,000 cards. A decision must be made on whether to send the cards out to be tabulated "as is" (accept the lot) or to check all of the 2,000 cards (reject the lot).

An acceptance sampling plan is used to make the decision. The following objectives are set for the decision procedure:

$$\begin{array}{ll} AQL: \pi_0 = .005 & \alpha = .05 \\ LTPD: \pi_1 = .025 & \beta = .10 \end{array}$$

The statistical decision rule to meet these objectives (assuming simple random sampling) can be found from the Table 8.1. The value of R is

$$\pi_1/\pi_0 = .025/.005 = 5.$$

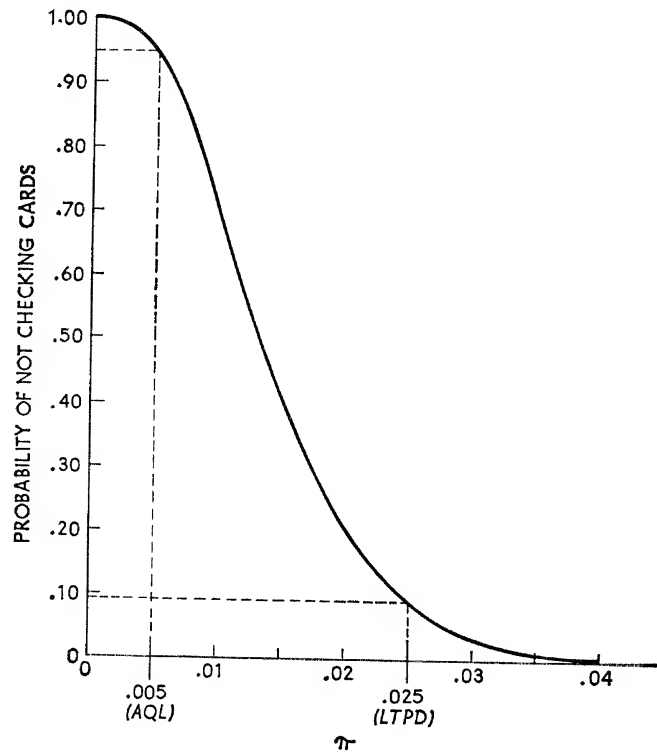
The next smaller value in the table is $R = 4.89$; c is 3, and $n\pi_0$ is 1.366. Therefore,

$$n = \frac{1.366}{.005} = 273,$$

or 275, for administrative convenience.

The statistical decision rule is: Select a simple random sample of 275 cards. Send the batch of 2,000 cards to be tabulated if there are 3 or fewer incorrectly punched cards in the sample; otherwise, check all of the 2,000 cards.

The operating characteristic curve for this decision rule appears as follows:



Note that at $\pi_0 = .005$ (the *AQL*) the probability of not checking the cards is approximately .95 (and of checking them is .05). At $\pi_1 = .025$ (the *LTPD*) the probability of not checking is approximately .10. This is as it must be, for we set the conditions, $\alpha = .05$ and $\beta = .10$, in stating the objectives of the sampling plan.

Suppose that we now review the basic principles of how to set π_0 , π_1 , α , and β . These principles have been implicit in our work, but let's make them quite explicit—even going so far as to number them:

1. In selecting values for α , β , π_0 , and π_1 , it usually is easiest first to set $\alpha = .05$ and $\beta = .10$ (or $\alpha = .01$ and $\beta = .05$, or some other pair of standard values). Then one should ask himself what value of π he wishes to associate with a .05 risk of rejecting that lot; the answer is the value of π_0 , or the *AQL*. Similarly one should ask himself what value of π he wishes to associate with a .10 risk of accepting that lot; the answer is the value of π_1 , or the *LTPD*.

2. In setting the *AQL* and the *LTPD*, as well as α and β , one must consider the consequences of possible decisions. The following questions must be considered: What are the consequences of accepting poor lots? What are the consequences of rejecting good lots? What is a poor lot? What is a good lot?

3. As a general rule, the acceptable quality level should be the quality level for which management is aiming. An acceptance sampling plan which has a good chance of *rejecting* lots of lesser quality may persuade vendors to ship lots of the acceptable quality.

4. The lot tolerance per cent defective set by management must not be an average quality or a quality level to which it is indifferent. *LTPD* should represent a relatively poor or barely tolerable quality level.

5. Both the *AQL* and the *LTPD* should be *possible* quality levels. For example, suppose that the lot tolerance per cent defective is set at .08 with a β -risk of .10. Then, if a lot as much as .08 defective is received, the chance of accepting it will be .10. However, if there is *no chance* that such a lot will be presented, the sampling plan is giving protection against something that will never occur! The plan is not the right one for the decision problem.

6. The values of π_0 and π_1 must not be too *close*. If they are, a large sample size will be necessary to achieve the specified risks. Often, in practice, after setting π_0 and π_1 management may find the required sample size uneconomical. The original goals may then be revised.

7. It must be recognized that the producer's risk (α), and the consumer's risk (β) are both risks to the decision maker. If a consumer rejects lots of good quality too often, he is committing the error associated with the producer's risk. However, these mistakes ultimately will be passed back to him in terms of bad will and higher prices.

8. If a series of lots are received over a period of time, it may be profitable to revise the sampling plan periodically.

8.6 Other Acceptance Sampling Plans

The acceptance sampling plans that we have been discussing are known as *fixed-sample size* or *single* sampling plans—a decision to accept or to reject a lot is reached on the basis of only one sample. At times, however, *double*, *multiple*, and *sequential* sampling plans are used.

Under a double sampling plan a first sample is drawn and one of the following actions is taken: (a) accept the lot, (b) reject the lot, or (c) take a second sample. If a second sample is selected, the alternatives then are to accept or to reject the lot. Thus, under a double sampling plan, an ultimate decision on the lot *may* be reached after the first sample is inspected but *must* be reached after the inspection of the second sample.

A multiple sampling plan simply is an extension of double sampling. Provision is made for the possibility of several samples. After each sample is inspected, a final decision on the lot is made or another sample is selected. However, an action to accept or reject the lot must be taken after a predetermined number of samples have been selected.

Under one type of sequential sampling plan, items are inspected one at a time. After each inspection a decision must be made to accept the lot, to reject the lot, or to select another item. In other sequential plans, sequences of groups of items rather than one item at a time are inspected.¹ In either event it is not possible to determine in advance the maximum number of items that must be sampled to reach a final decision on the lot. At times, therefore, a sequential plan may be truncated—i.e., a decision must be reached after a predetermined number of samples. Such truncated plans simply are a type of multiple sampling.

For a given π_0 , π_1 , α -risk, and β -risk, the single, double, multiple, and sequential sampling plans can be determined by reference to tables which are similar to Table 8.1 and which have been developed by statisticians. Each plan so found provides approximately the same protection against incorrect actions. That is, the operating characteristic curves for the four sampling plans are almost identical.

¹The theory underlying sequential sampling plans was first developed in 1943 by Abraham Wald (1902–50), a mathematical statistician who was responsible for many important contributions to modern statistics.

Table 8.2 presents similar single, double, and multiple sampling plans. For each plan the *AQL* is 4 per cent and the *LTPD* is 9 per cent. Also, in all cases, $\alpha = .05$ and $\beta = .10$. For each sample that might be selected, an acceptance number and rejection number is listed. If the number of defectives is equal to or less than the acceptance number, the lot is accepted. If the number of defectives is equal to or larger than the rejection number, the lot is rejected. If the number of defectives is greater than the acceptance number

Table 8.2

SINGLE, DOUBLE, AND MULTIPLE SAMPLING PLANS

$$\begin{array}{ll} \pi_0 = .04 & \pi_1 = .09 \\ \alpha = .05 & \beta = .10 \end{array}$$

Type of Sampling	Sample	Sample Size	Combined Samples		
			Size	Acceptance Number	Rejection Number
Single	First	225	225	14	15
Double	First	150	150	9	24
	Second	300	450	23	24
Multiple	First	50	50	1	6
	Second	50	100	3	9
	Third	50	150	7	13
	Fourth	50	200	10	16
	Fifth	50	250	13	19
	Sixth	50	300	16	22
	Seventh	50	350	19	25
	Eighth	50	400	24	25

By permission from Columbia Statistical Research Group, *Sampling Inspection* (New York: McGraw-Hill Book Co., Inc., 1948), p. 330.

but less than the rejection number, another sample is selected. Both the acceptance and the rejection numbers refer to the number of defectives in the combined samples. Thus, for example, consider these numbers for the fifth sample of the multiple sampling plan. If a fifth sample need be drawn and if 13 or less of the 250 items selected for the five samples are defective, accept the lot. If 19 or more of the 250 items are defective, reject the lot. If from 14 to 18 items are defective, select a sixth sample.

Note that under double sampling an ultimate decision on the lot must be reached after the second sample is inspected. With the multiple sampling plan under consideration, an ultimate decision must be reached after the eighth sample is inspected. This is apparent for the rejection number is one larger than the acceptance number.

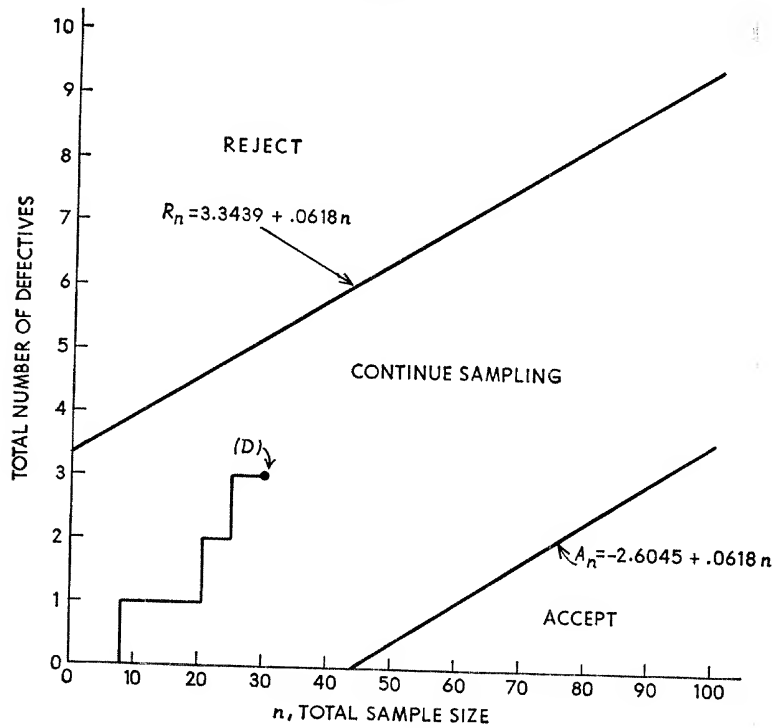
When sequential sampling is employed, it often is simpler to present the decision rule as a chart. Figure 8.3 presents a sequential sampling plan based on the same π_0 , π_1 , α -risk, and β -risk as the plans presented in Table 8.2. The acceptance and rejection numbers are given by the following equations:

$$A_n = -2.6045 + .0618n$$

$$R_n = 3.3439 + .0618n.$$

A_n and R_n are the acceptance and rejection numbers, respectively, for a cumulative sample of n items. These lines divide Figure 8.3 into three regions: an acceptance region, a rejection region, and a

Figure 8.3

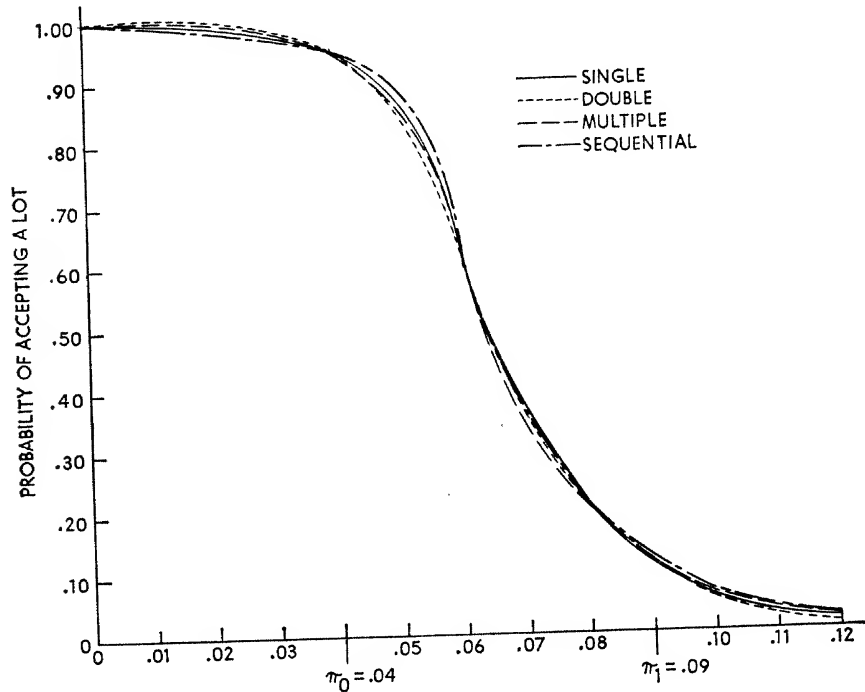


region for further sampling. After each item is inspected, a point is plotted at a height equal to the combined number of defectives. For example, one point (D) is marked on the chart. This point indicates that after the thirtieth item is inspected, 3 of the 30 items are defective. Since this point falls between the two lines, sampling is continued. If a plotted point falls on or above the

upper line, the lot is rejected. If a point falls on or below the lower line, the lot is accepted.

Note that under this plan, a lot cannot be accepted until more than forty (43) items have been inspected. This is the first time that it is possible for a point to fall into the acceptance region. This, however, is less than the smallest sample size, 50, required to reach a decision under the equivalent multiple sampling plan. It is much

Figure 8.4



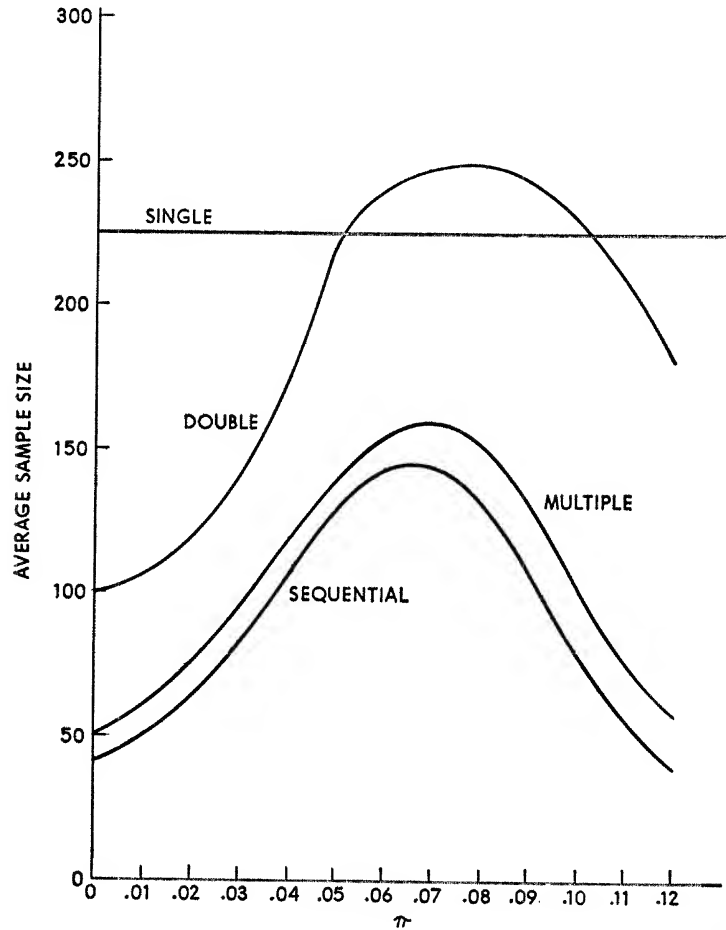
smaller than the minimum sample size for double and single sampling. Note also that with sequential sampling lots can be rejected with samples as small as size 4. This is far smaller than the minimum size of the required samples for the other plans.

In Figure 8.4 we present the OC curve for the four sampling plans. As we had anticipated, they all are practically identical. Thus, so far as protection against incorrect decisions is concerned, they are interchangeable.

The choice of one plan as against another depends on such practical considerations as the costs and ease of inspection and administration. The chief advantage of double, multiple, and sequential

sampling plans is that they may require a smaller sample size, on the average, to reach a decision on a lot. For a single sampling plan, the sample size is fixed no matter what the quality of an incoming lot may be. Under the other plans, the number of items needed to reach a decision varies. It is possible, therefore, to reach a decision under

Figure 8.5



these plans with a smaller average sample size than under single sampling plans. In particular such economies will occur whenever incoming lots are either of very good or of very poor quality. A small sample, on the average, will be sufficient to detect such extremes in quality. For lots of intermediate quality, larger samples may be required to reach an ultimate decision. In Figure 8.5 we present curves which depict the relationship between the average sampling num-

ber and the quality of incoming lots for the four sampling plans considered in this section. The relationships presented are fairly typical. Sequential sampling uses the smallest average sample size. Generally speaking, multiple, double, and single sampling follow in that order. Note, however, that for lots of a certain quality (from 5 to 10 per cent defective) the average sample size for the single sampling plan is less than for double sampling.

A second advantage claimed for double, multiple, and sequential sampling plans is their psychological appeal. Suppliers often feel that such plans are more equitable since more than one chance is given a lot before a possibly adverse decision is made. Such beliefs, however, are erroneous for actually vendors are given the same protection under any plans for which the *OC* curves are alike.

Shortcomings of double, multiple, and sequential plans arise from the administrative difficulties they present. They involve more record keeping than single sampling plans. Hence, it is more difficult to train inspectors, and it increases the likelihood of clerical errors. Under single sampling plans the sample size per lot is constant, and, hence, it is easy to set up inspection schedules. Under the other plans sample size varies from lot to lot, and inspectors may be overworked at times and underworked at other times. Finally, for double, multiple, and sequential plans physical problems of selecting more than one sample may exist. Thus, if all the samples that might be needed under a multiple sampling plan are drawn before inspection begins, too many items might be selected and costs incurred needlessly. However, if samples are drawn as needed, the inspection department might become cluttered with lots in the process of inspection. In general, then, the pertinent facts surrounding a given inspection situation will dictate the appropriate method of sampling.

8.7 Management Objectives and Decision Rules

In the previous sections of this chapter we considered the determination of decision rules to provide management with the requisite degree of protection against the two types of possible errors in two-action problems. Management sets its goals in terms of an α -risk associated with one state of the world and a β -risk associated with a second state of the world. Statistical theory then enables us to determine the sample size and the critical value of the statistic.

However, in setting its objectives in this way, management gives

up its control over the sample size—an important factor in the cost of a sampling procedure. Once the risks are set, the sample size follows directly from the statistical techniques employed. The sample size obtained in this manner gives management the desired protection but at times its cost may be prohibitive. Then, again, the size of the sample may be so large that the decision procedure cannot be employed expeditiously. Thus, consider a situation where it is necessary to check a production process at regular intervals to decide whether or not the process is operating satisfactorily. If the inspection process is a lengthy one—for example, involving laboratory tests—a large sample may be too time consuming. The damage may be done before the sample results are made available.

Under such circumstances it may be desirable to set the sample size in advance. However, in this case, we can control the probability of only one type of error. In other words, for the advantage of being able to control the cost or expeditiousness of a sampling procedure, the control over one type of error is sacrificed. Therefore, we must look for a decision rule which will make use of the desired sample size and give us the desired protection against the type of error we are more eager to avoid. Subsequent to the determination of the decision rule, risks of other errors also may be computed.

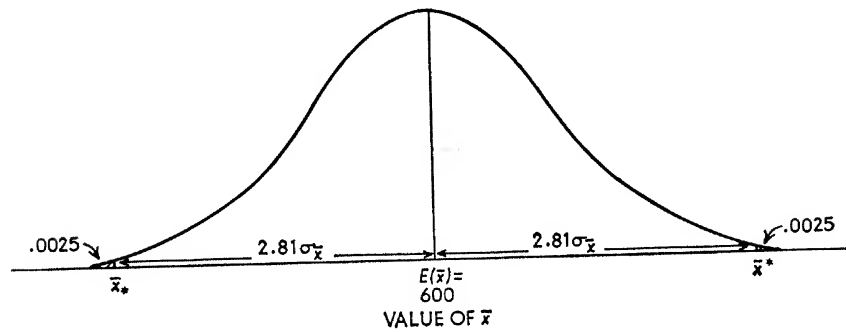
To illustrate the determination of a decision rule under such circumstances, consider the following example.

Example 8.6 A company samples 4 steel rods every half-hour and determines their average tensile strength. On this basis, it decides whether a process is or is not operating satisfactorily. Past experience has indicated that a desirable process average is 600 pounds per square inch, and the process standard deviation is 20 pounds per square inch. It is willing to run a risk of .005 of stopping the process when it is operating satisfactorily. Thus, the company is more eager to avoid the risk of stopping the process unnecessarily rather than leaving the process alone if it is not operating satisfactorily.

The objectives that the decision rule is to attain are $n = 4$ and $\alpha = .005$. The rule must specify the critical values of $\bar{x}$. The sampling distribution of $\bar{x}$ if $\mu = 600$ will be approximated by a normal distribution with $E(\bar{x}) = 600$. The standard error of the mean is

$$\begin{aligned}\sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \\ &= \frac{20}{\sqrt{4}} \\ &= 10.\end{aligned}$$

The sampling distribution and the area conforming to an α -risk of .005 are shown in the diagram below:



The diagram indicates that

$$\bar{x}_* = \mu - z\sigma_{\bar{x}} \quad \text{and} \quad \bar{x}^* = \mu + z\sigma_{\bar{x}}.$$

Since the area in each tail is .0025, $z = 2.81$. Hence

$$\begin{aligned} \bar{x}_* &= 600 - (2.81)(10) & \text{and} & & \bar{x}^* &= 600 + (2.81)(10) \\ &= 572 & & & &= 628. \end{aligned}$$

The decision rule is: Select a simple random sample of 4 items. If $\bar{x}$ is between 572 and 628, leave the process alone. If $\bar{x}$ is equal to 572 or less, or if $\bar{x}$ is equal to 628 or more, stop the process. We could determine the *OC* curve of this decision rule if it were desirable to do so. This would indicate the risks of incorrect decisions not only for $\mu = 600$ but for all other possible values of μ .

There are, however, other goals which management might set for a decision rule. Consider an acceptance sampling situation where the alternative courses of action are to inspect and not to inspect the lot. Under such circumstances, management may require a decision rule that would keep total inspection costs of good lots at a minimum. The total cost includes the cost of sample inspection and the cost of inspecting lots that are not accepted. But, to obtain this objective, management must again give up control over the risk of committing one type of error. We shall not discuss the development of decision rules when the objectives of a decision rule are set in this way. However, these objectives are very common, and tables have been worked out to facilitate the determination of decision rules. There are tables to provide minimum inspection of good lots and control over the probability of accepting poor lots. You will find references covering such tables and their use at the end of the chapter.

We have not listed all the goals management might set for a decision rule. You are aware, however, that there are several. Further-

more, in two-action problems, it is possible to control only two independent objectives. In setting goals it is, therefore, most important to determine which objectives it is most economic to control.

8.8 You Should Now Know That

A statistical testing problem involves a choice between two or more qualitative courses of action—the decision to be based on a sample of observations.

To treat a testing problem statistically states of the world must be described by a decision-parameter.

Two-action problems are those in which only two alternative courses of action are available.

These problems present the possibility of two types of incorrect decisions: called errors of the first kind (or Type I errors) and errors of the second kind (or Type II errors).

By convention, a problem should be formulated so that the more serious error is considered the error of the first kind.

The probability of an error of the first kind is called the α -risk. The probability of an error of the second kind is called the β -risk.

Both of these risks are conditional upon the existence of a particular state of the world (a particular value of a decision-parameter).

It is possible to describe decisions in terms of the acceptance and the rejection of hypotheses.

A hypothesis is an assumption about which state of the world is true.

A decision rule for a testing problem must specify (a) the method of sampling, (b) the sample size, and (c) critical values which relate all possible values of a statistic to decisions.

Decision rules may be one-tailed or two-tailed, depending on the nature of the problem.

One method of determining statistical decision rules for testing problems requires one to focus attention on pertinent states of the world, to specify α - and β -risks, and to solve appropriate equations for the sample size and critical value(s).

An *OC* curve depicts, for all possible values of a decision-parameter, the risks of relying on a given decision rule.

The larger the sample size, the steeper is the *OC* curve.

Acceptance sampling is the application of statistical testing to decision problems in industrial and clerical inspection.

In acceptance sampling the α -risk is called the producer's risk; the β -risk is called the consumer's risk.

The two values of the decision-parameter which describe states of the world in acceptance sampling often are called the *AQL*, acceptable quality level, and the *LTPD*, lot tolerance per cent defective.

At times it is economical to use double, multiple, or sequential sampling plans in acceptance/rejection decision problems.

In the determination of decision rules, it is possible to control only two independent objectives. In setting goals, it is, therefore, important to determine which objectives it is most economic to control.

8.9 You Should Now Be Able to Solve

1. Would you or would you not suggest a statistical testing procedure in each of the following problem situations? Explain your answer.

- a) An auditor must decide whether to approve a payroll for 5,000 workers or to check the entire payroll for clerical errors.
- b) A cost accountant wants to determine the average waste produced in carrying out the operations of a given department as a basis for deciding on whether or not to recommend revamping those operations.
- c) A company wants to decide whether or not to expand the size of its plant.
- d) A company wants to take remedial action if the proportion of misfiled papers of its filing department increases.
- e) Management must make a decision on whether or not to promote an assistant buyer to the rank of buyer in a department store.

2. A company receives glass tubes in lots of 10,000. The proportion of broken tubes, π , is considered the decision-parameter. It desires an acceptance sampling procedure that will offer it the following protection: If $\pi = .004$, the probability of rejecting the lot is to be .05 (α -risk); if $\pi = .020$, the probability of accepting the lot is to be .10 (β -risk). Use Table 8.1 (p. 259) to determine the decision rule.

3. A firm that employs 2,500 people wishes to control excessive lateness. Past experience has indicated that, on the average, only 2 per cent of its employees are late. The firm wishes to maintain this average. The firm desires a statistical decision rule which will enable it to decide when and when not to take action on excessive tardiness.

- a) What are the two types of incorrect action that might be taken?
 - b) Is the population of interest finite or infinite?
 - c) What is an appropriate decision-parameter?
 - d) The firm decides to use $n = 2,500$ (all of the employees). Is the problem still a sampling problem? Explain.
-

- e) Since n is given, only one risk can be specified. Which of the two types of error would you desire to control? Explain.
- f) Set a risk, that you are willing to take, of making the type of error that you have selected in (e). What factors did you take into consideration in setting this risk?
- g) Is the decision rule one-tailed or two-tailed?
- h) Derive the decision rule.
- i) Determine and graph the OC curve for this rule. What does this curve show?

4. A large grocery chain takes four inventories of canned goods a year in each of its stores. If there is a shortage or overage between the physical check and the record of the store's inventory in the central office, further investigation is to be made. Management seeks a statistical decision rule to relate observations to the appropriate action. The decision-parameter that is agreed upon is the population mean—the average value per physical unit of inventory. Assume that $\sigma = 15$ cents.

The following states of the world and risks of incorrect decision are decided upon:

$$\begin{aligned}\mu_0 &= \text{average value based on the company's records} \\ \mu_1 &= \mu_0 - \$0.01 \\ \mu_2 &= \mu_0 + \$0.01 \\ \alpha &= .01 \\ \beta &= .05\end{aligned}$$

- a) What is the Type I error in this problem?
 - b) What is the Type II error?
 - c) Explain the meaning of $\alpha = .01$.
 - d) Explain the meaning of $\beta = .05$.
 - e) Is the decision rule for this problem one-tailed or two-tailed? Explain.
 - f) What values must be computed to determine the desired decision rule?
 - g) Derive the decision rule (assume μ_0 equal to any reasonable value).
 - h) Determine and graph the OC curve for the decision rule. How do you interpret the meaning of this curve?
5. A company wants to decide on whether to use iridescent wrappers to package its soaps or not. You are asked to formulate a statistical decision rule.
- a) What would be an appropriate population and an appropriate decision-parameter? Explain.
 - b) Should the rule be a one-tailed or a two-tailed rule? Explain.
 - c) Select values of the decision-parameter upon which to focus attention. Explain the reasons for your choice.

- d) What are the possible types of incorrect action in this situation? Explain.
- e) Set what you consider economical risks of committing each error. Explain the reasons for your choice.
- f) Which is the α -risk? Which is the β -risk? Explain.
- g) Determine the decision rule.
- h) Would it be helpful to determine the OC curve for this decision rule? Why or why not?

8.10 You Will Also Find That

Key general references on statistical testing, hypotheses, and operating characteristic curves include:

NEYMAN, J. *First Course in Probability and Statistics*, pp. 250–67. New York: Henry Holt & Co., Inc., 1950.

SPROWLS, R. CLAY. *Elementary Statistics*, pp. 51–95, 183–206. New York: McGraw-Hill Book Co., Inc., 1955.

For discussions of the applications of statistical testing to industrial inspection decision problems (acceptance sampling), the following references are useful:

GRANT, EUGENE L. *Statistical Quality Control*, pp. 311–38. 2d ed. New York: McGraw-Hill Book Co., Inc., 1952.

DUNCAN, ACHESON J. *Quality Control and Industrial Statistics*, pp. 131–60. Rev. ed. Homewood, Ill.: Richard D. Irwin, Inc., 1959.

A survey of acceptance sampling procedures for use in accounting and in clerical operations is contained in:

VANCE, LAWRENCE L., and NETER, JOHN. *Statistical Sampling for Auditors and Accountants*, pp. 27–52, 91–101. New York: John Wiley & Sons, Inc., 1956.

CHAPTER • 9

Error and Bias

9.1 Introduction: Sampling Errors versus Procedural Bias

Not many years ago one well-known statistician pointed out that in the teaching and the learning of statistics “perhaps the fundamental thing is to bring the students to the point where they will never again look at a number with the same naïve, trusting confidence they had as grade school children.”¹ This chapter focuses on this idea.

In the past two chapters we have considered risks—risks of incorrect decisions in estimation and in testing problems. These risks are the risks of acting on the basis of partial or sample rather than complete information. They are due to the possibility of sampling error—the difference between a sample result and the result that would have been derived from a complete count. It is important to remember that statistical theory alone provides procedures for selecting samples so that one may measure and control the risks due to sampling.

However, it is equally important to recognize that in addition to sampling errors there are many other types of possible “errors” that may mar the results of a statistical study. These also present risks of making incorrect decisions to a decision maker. These other types of errors are not due to sampling. They would arise even though a complete count were taken. They are biases which result either from the way the problem was defined or from an imperfection in the method used to elicit information. Hence, they are called *procedural biases* or *nonsampling errors*.

¹ Irving W. Burr, “What Principles and Applications of Statistics Should Be Taught in High School and Junior College?” *The Mathematics Teacher*, Vol. XLIV (1951), p. 12.

For example, suppose that a decision problem is incorrectly formulated—say, by specifying the wrong decision-parameter. This may result in an incorrect decision—it certainly adds to the risks of taking the wrong action. This imperfection is not, however, an error of sampling. It is a built-in deficiency. It may bias the results of a sample, but it also may bias the results of a complete count.

Consider, also, a marketing research study of incomes in a community. If observations are to be made by requesting families to provide this information about themselves, the possibility of bias exists. Persons who would tend to overstate or to understate their incomes might introduce a bias into the results of the study—and this would be true whether a sample or a complete count were taken.

The results of almost every statistical study are subject to some imperfections—sampling errors, procedural biases, or, more usually, both. We have studied, in the preceding chapters, how statisticians have developed procedures which allow the measurement and the control of risks of sampling error. In this chapter we will study procedures for the control of bias.

Imperfections in numerical information, whether they be the result of sampling or of a biased procedure, are of obvious importance to a decision maker. He cares not why an incorrect decision is made—to him only the fact that erroneous information led him astray is important. For example, what would be the value of a carefully and properly designed probability sample for a marketing research study if the questionnaire was so poorly designed that it elicited biased and, hence, useless answers? Thus, in designing a statistical decision procedure we must consider not only the risks of relying on sample information but also the risks of relying on information that may be biased.

Nevertheless, the distinction between sampling error and procedural bias is important, and in speaking of the degree of the trustworthiness of data from any study, statisticians use some special terminology to point up the distinction. The *precision* of a procedure is measured by the size of the potential sampling error. The *accuracy* of a procedure refers to the magnitude of the sampling error and the potential bias in information. Thus, to point up this distinction, note that data collected by a complete count are perfectly precise, but they are not necessarily very accurate. Their accuracy depends on the extent and the magnitude of procedural biases.

Unfortunately, the control of biases is operationally limited because few special techniques for this task have been developed. In

contrast to sampling error, there are almost no known methods of measuring the magnitude of biases, although some statistical techniques are available to check for the presence of such biases. In any event, management and the statistician always must be alert to the possibility of both error and bias. They can control the former by using probability sampling. However, generally speaking, the only way of controlling biases is by so planning a study that the opportunities for their occurrence are limited. In order to do this, those conducting a sample study or a complete count must be aware of the various opportunities for bias. Once aware of these opportunities they may prepare for them.

Thus, in the subsequent sections of this chapter we shall survey various types of procedural biases and see how in primary studies, proper planning can minimize the opportunities for their occurrence. Knowledge of the ways in which biases may occur also is important to a decision maker who is using secondary data—information collected for another purpose, perhaps under some other auspices. Without knowledge of the accuracy of such data, how can one have confidence in them? One should be suspicious of almost all data until he verifies whether or not they were collected by a reliable and valid *procedure*. Of course, some data are more trustworthy than others, and the auspices under which data are collected often is a guide to their accuracy. One need not reject the trustworthiness of all numerical information. He merely should be suspicious of almost all data—at least until he is familiar enough with the procedure by which they were collected.

9.2 Improper Formulation of the Decision Problem

No type of bias is more unfortunate than that which results from the improper formulation of the decision problem, *per se*. Such imperfections often result in a statistical study providing the right answers to the wrong questions—which only can mean the right decision for the wrong decision problem. Unfortunately, there are several specific opportunities for such deficiencies to be built into any study.

First of all, bias may result from a failure to think through and specify carefully enough the essential ingredients of the decision problem (the alternative courses of action, states of the world, and consequences). Failing to recognize exactly what actions can be taken and upon what factors these actions should depend can lead only to the wrong definitions of the population and of the decision-

parameter(s). And as we know, no sample (nor any complete count) can provide information about anything but the population from which it was selected.

Example 9.1 To point up the fact that it is essential that any statistical population of study be defined properly—that it be defined by management and the statistician to include pertinent observations—one statistician once pointed out the following incident:

“Take the case of the manufacturer of dog food . . . who went out and did an intensive market study. He tested the demand for dog food, he tested the package size, the design, the whole advertising program. Then he launched the product with a big campaign, got the proper distribution channels, put it on the market and had tremendous sales. But two months later, the bottom dropped out—no follow-up sales. So he called in an expert, who took the dog food out to the local pound, put it in front of the dogs—and they would not touch it. For all the big premarketing study, no one had tried the product on the dogs.”²

Thus, the population of study must be specified carefully. The definition of an elementary unit must be clear; that of an observation unambiguous. Unfortunately, this is not always easy. Consider the problem of estimating total unemployment in the United States on a given date. The number of unemployed will depend on the definition given “an unemployed person.” Unfortunately, economists differ in their opinions of what is an acceptable definition. What one may consider a valid definition, another may regard as improper. Hence, one economist will regard a study as unbiased; another will consider it biased. Needless to say, statistics cannot solve this problem; the answer, if there is one, must be provided by economics.

An additional problem in some studies is that a satisfactory frame which identifies the population must be provided. The problem of securing an adequate frame is not always simple and can be costly—as we saw in our discussion of frames. However, the cost of a good frame must always be weighed against the defects in a bad one. We have considered earlier how frames may suffer from defects—incompleteness, duplication, inadequacy, inaccuracy, etc. These defects may result in the data collected being seriously biased. Why? The frame fails to identify the correct population with a reasonable degree of accuracy.

² Joseph R. Hochstim, “Practical Uses of Sampling Surveys in the Field of Labor Relations,” *Proceedings of the Conference on Business Application of Statistical Sampling Methods* (Monticello, Ill.: The Bureau of Business Management, University of Illinois, 1950), pp. 181–82.

Example 9.2 Suppose that the school board of a community wishes to determine the attitudes of families with children in school towards a suggested modification in their schools' extracurricular programs. A readily available frame is a list of school children. However, this frame will duplicate the names of several families for some families will have two or three or more children in school.

Thus, the available frame suffers from a defect—duplication—which if allowed to stand might result in a bias in the results of the study. The potential bias arises from the fact that the frame affords some families, perhaps with atypical attitudes, a greater chance to appear in, say, a simple random sample from the frame.

One possible solution is to ignore this bias. Another, obviously more costly but often more preferable alternative, is to check the frame and sort out the duplicates. In addition, some sampling techniques have been developed so that a sample may be drawn and estimates prepared in such a way that the bias is minimized.

Also note that the suggested frame would be incomplete if one were trying to study the attitudes of all families in a city—families without school children would not be identified by the frame. This would, more or less, bias the results of a study, depending on its nature and purpose.

As our discussion in this section indicates, biased information may result not from the failure of statistical sampling but from the exercise of poor judgment in the formulation of a decision problem. This is why many persons have pointed out that management, as well as a statistician, must take an active role in the planning of any study. Together, these persons, aware of the opportunities for bias due to the improper formulation of the decision problem and due to other factors as well, may so plan a study as to minimize the potential effects of bias.

9.3 Biased Selection Procedures

In discussing methods of probability sampling you will recall that in all cases we defined a particular method in terms of the probabilities of selection it should yield. For example, simple random sampling was defined as a method of selection which gives every possible sample of size n the same chance of being the sample chosen in a study. Subsequent to these definitions, we considered administratively feasible procedures for selecting samples. Thus, we used random numbers to draw simple random and other probability samples. Implicit in the use of any such procedure is the assumption that it meets the probabilistic definition of the method adopted. Thus, for example, if we use random numbers in order to select a

simple random sample, we are assuming that if the random numbers were used repeatedly every possible sample of size n would occur approximately the same number of times.

Unless our *modus operandi* meets the theoretical definition of the specified method of sampling, we say that our sampling procedure is biased. If it does not give, for example, every sample combination of size n the same chance of being selected when it should, we have a biased procedure for selecting a simple random sample. A biased sampling procedure is undesirable because it means that our calculations of sampling risks are based on an erroneous assumption, which leads us to think that we know these risks when actually we do not.

Example 9.3 Consider the problem of checking on the quality of a filing process in order to provide a basis for a decision on whether or not to install a modern records system. Assume a simple random or some other type of probability sample is to be drawn from existing records and that the records are filed in floor-to-ceiling cabinets.

If a sample were to be drawn from this system without the benefit of random numbers, several opportunities for a biased sampling procedure exist. Thus, if clerks were asked to pick a random sample from the files, there might be an unwitting tendency to avoid the selection of records in the highest or in the lowest cabinets. Those records would be less conveniently available and perhaps their chances of appearing understated. But in the records-keeping operation the more inconvenient position of these cabinets may result in more filing errors. For a sampling procedure to be unbiased these "high-low" records should be given their proper chance to appear in a sample.

To eliminate this aforementioned bias, and any others which we cannot even suspect, it would be best to use random numbers to determine the elementary units selected for the sample.

Thus a way to avoid bias in a sampling procedure is to use random numbers. Is this a good solution? Statisticians are able to answer "yes" on the basis of certain tests that they apply to random numbers and on the basis of their experiences. The repeated use of such tables has indicated that they bring one very, very close to satisfying the various theoretical definitions of probability sampling methods.

It is well always to remember the danger of bias with judgment or convenience sampling. In these cases probabilities of selection cannot be ascertained. In fact, probability has no meaning if judgment or convenience rather than a random process is the basis for selection. The result is that there is no known way of generally evaluating judgment or convenience sampling. Furthermore, many experiences where a later test was available have shown that these

methods have led to inaccurate results—due to errors of judgment, selectivity, and availability. In this connection it is worthwhile to recall Examples 4.14, 4.15, and 4.16 of Chapter 4.

The few last comments in this section refer to errors in the interpretation of the results of selection procedures, and to errors in the measurement of risks. One refers to the bias of choosing among several samples. For example, suppose that we were to flip a fair coin 5 times. The result would be a simple random sample of 5 outcomes. The probability of no tails in one such sample would be small—only .0312. However, if *several* samples of 5 were derived, the chance of no tails in one or more samples would increase significantly. This probability would be .1468 for five samples; .2721 for ten samples; and .5479 for twenty-five samples. By repeating any sampling procedure often enough, we can make even the rarest of outcomes very likely to occur at least once. Of course, if in twenty-five samples of 5 flips of a coin we did observe zero tails once or twice, we hardly would assert that we had proved tails never occur. However, some persons in doing an experiment repeatedly will ignore several unfavorable results and become astounded and overly impressed with one favorable result. Suppose that we quote a hypothetical example.

Example 9.4 "Users report 23 per cent fewer cavities with Doakes' tooth paste, the big type says. You could do with twenty-three per cent fewer aches so you read on. These results, you find, come from a reassuringly 'independent' laboratory, and the account is certified by a certified public accountant. What more do you want?"

"... But let's get back to how easy it is for Doakes to get a headline without a falsehood in it and everything certified at that. Let any small group of persons keep count of cavities for six months, then switch to Doakes'. One of three things is bound to happen: distinctly more cavities, distinctly fewer, or about the same number. If the first or last of these possibilities occurs, Doakes & Company files the figures (well out of sight somewhere) and tries again. Sooner or later, by the operation of chance, a test group is going to show a big improvement worthy of a headline and perhaps a whole advertising campaign. This will happen whether they adopt Doakes' or baking soda or just keep on using their same old dentifrice."³

Deliberate or not, bias due simply to picking among samples occurs more frequently than you may imagine. It affords another reason for checking one's decision procedure carefully and analyzing it logically.

³ Darrell Huff, *How to Lie with Statistics* (New York: W. W. Norton & Co., Inc., 1954), pp. 37-38.

A somewhat similar bias may occur in acceptance sampling.

Example 9.5 Acceptance sampling plans, as we know, are used to decide between accepting and rejecting a given lot. Suppose that rejected lots are returned to the producer. A serious problem is created if the producer chooses to resubmit the rejected lot without attempting to improve its quality.

Thus, suppose that for a given decision rule the risk of accepting a lot of mediocre quality is only .20. This is the probability of accepting it on one trial. If a rejected lot also is resubmitted, say, two other times, this probability more than doubles since:

$$\begin{aligned}
 P(\text{ACCEPTANCE ON FIRST SUBMISSION}) &= .200 \\
 P(\text{REJECTION ON FIRST SUBMISSION AND} \\
 &\quad \text{ACCEPTANCE ON SECOND}) = .80 \times .20 = .160 \\
 P(\text{REJECTION ON FIRST TWO SUBMISSIONS} \\
 &\quad \text{AND ACCEPTANCE ON THIRD}) = .80 \times .80 \times .20 = .128 \\
 P(\text{ACCEPTANCE ON ONE OF THREE SUBMISSIONS}) &= .488
 \end{aligned}$$

The consumer's risk thus increases rapidly if rejected lots are resubmitted. Hence, the original risk of .20 becomes meaningless. Several solutions are possible. One of the more obvious may be noted: If possible, identify and fully check resubmitted lots of product.

One other type of bias in the interpretation of the results of sampling procedures often is caused by the failure to recognize that risks of committing errors always exist when samples are the basis for a decision. Again we quote on the subject of a dentifrice.

Example 9.6 "Something called Dr. Cornish's Tooth Powder came onto the market a few years ago with a claim to have shown 'considerable success in correction of . . . dental cavities.' The idea was that the powder contained urea, which laboratory work was supposed to have demonstrated to be valuable for the purpose. The pointlessness of this was that the experimental work had been purely preliminary and had been done on precisely six cases . . ."⁴

The poor decision illustrated by this last example rests with Dr. Cornish's failure to consider the risks associated with a sample of but six cases. The risk of marketing the product was obviously great.

9.4 Bias of Nonresponse

In our previous discussions we have assumed that once a sample of n elementary units is planned, it is always possible to observe the pertinent characteristics of those particular elementary units. That

⁴ *Ibid.*, p. 38.

is, we have assumed only the possibility of a 100 per cent response to any sample. The failure of this assumption is called nonresponse.

Some nonresponse will occur in most statistical studies. It is particularly common if a study involves people as elementary units. Thus, nonresponse may arise in such studies from (a) people not found at home—perhaps moved, or even deceased; (b) a person's telephone being busy, disconnected, or remaining unanswered; (c) refusals of persons to give information; (d) peoples' inability to provide the pertinent information; and/or (e) people's failure to take the trouble to respond to a questionnaire. Analogous causes result in nonresponse in other types of statistical studies. To consider a few causes, note the possibilities of lost, destroyed, or misfiled records, outstanding checks, a lot of misplaced parts, library books in a bindery and not immediately available, etc.

Nonresponse may impair the usefulness of a statistical study. The main difficulty rests with the fact that respondents may differ from nonrespondents with respect to the characteristic of interest. The result—the respondents in a study provide only biased information about the population as a whole. The study suffers from the *bias of nonresponse*.

Example 9.7 In one application of statistics nonresponse reflected itself in the following way:

"A sample of 208 manholes belonging to a large utility company appealing to the court for higher rates was drawn for the purpose of inspecting the manholes themselves, the ducts and the cables therein. It was then discovered that 22 of the manholes designated for the sample had been paved over. Civic authorities would not let them be uncovered. Substitutions could not be permitted, nor a new sample. The difficulty was settled unfavorably to the company by declaring them to be one grade lower than average."⁵

In this case, as you can see, if all paved-over manholes were discarded from the sample, that part of the population would have been given no chance to appear in the sample.

It is important to recognize that nonresponse may bias the results of an attempted complete count as well as those of a sample. In a complete count there will be more respondents, but there also will be more nonrespondents, and the relative frequency of them will be about the same as if a sample (probability, judgment, or convenience) were taken. Any differences in the characteristic of interest between the two groups will lead to biased results no matter what the

⁵ Reprinted with permission from W. Edwards Deming, *Some Theory of Sampling* (New York: John Wiley & Sons, Inc., 1950), pp. 13-14.

design of a study. Thus, nonresponse is an equally important consideration in all types of studies.

Example 9.8 In 1937 the United States Congress appropriated funds for a "census of unemployment." Although a complete count was aimed for, provision was made only for voluntary responses. To increase the amount of response, the President appealed for co-operation, a publicity campaign was conducted, and the Post Office Department undertook to leave a questionnaire at almost every doorstep in the country.

The rate of response was high, at least for this type of survey. A subsequent probability sample for which nonresponse was minimized was taken in order to evaluate the census returns. It indicated that about two thirds of those who should have responded to the census actually did so. However, it also indicated that there were several significant and substantial biases, due in part to nonresponse, in the (attempted) complete count.

The possible bias from nonresponse is sometimes ignored, and the respondents to a statistical study are considered a sufficient sample. But what type of sample do they represent? The answer is "a convenience sample." As we know, convenience sampling has its applications. However, providing a basis for an important decision is not one of them. Thus, the interpretation of respondents as a sample is particularly dangerous when that sample is blindly regarded as "representative."

One amazingly poor solution to nonresponse is sometimes suggested—increase the original sample size. Thus, suppose that a sample of 100 student answers to a questionnaire was desired and it were known, on the basis of past experience, that only about 25 per cent of the students contacted would respond. Should a sample of 400 be drawn on the basis of the arithmetical calculation: $400 \times 25\% = 100$? The answer is no—increased sample size will result in more respondents *and* more nonrespondents. Hence, it fails, in itself, to decrease any bias of nonresponse. It seems that the inadequate solution of increased sample size often is based on the erroneous idea that any large sample is a good one. (You may reconsider Example 4.16 as an aid in debunking this thought.)

What can be done about nonresponse and the bias it may interject into the results of a statistical study? The first general answer is "try, in so far as is possible, to prevent it by a well-planned procedure for obtaining information."

Example 9.9 The settlement of an excise tax case required that the inventory of the taxpayer, who operates several variety stores, be

determined. The examining agent of the Internal Revenue Service suggested that the inventory on a given date be taken so as to include brand names. Instructions to clerks who were to take the inventory included this requirement. However, due to poor supervision, brands often were not listed on the inventory record sheets. This was not discovered until a later date.

Thus, in a sense the study suffered from nonresponse—some pertinent observations, though supposedly drawn into the sample, were missing. The date of the inventory was far in the past, and there was no way available to remove the nonresponse. If brands were actually pertinent, a better inventory-taking procedure would have eliminated the nonresponse.

The proportion of nonresponse in any study will depend, in part, on controllable factors. Thus, in the sampling of human populations, the questionnaire, the interviewers, and the publicity are among the determinants of the amount of nonresponse. For example, it has been observed, in many studies, that some interviewers have persistently good luck at inducing people to respond, while others have persistently bad luck.

Few statistical studies will be completely free of nonresponse even with the best-planned procedures. However, the proportion of nonresponse in a study may be minimized further by allowing for *callbacks* or *recalls*. Some nonrespondents may be turned into respondents by supplementary contact. Two, three, four, or more appeals often are used to attempt to eliminate a good part of the original set of nonrespondents. Letters, telephone calls, and personal interviews may be used, for example, to supplement an original questionnaire. Furthermore, a sample of nonrespondents may make the provision for recalls in a survey more feasible.

Example 9.10 In one statistical study to determine interest in a radio program on the topic of child training, an original questionnaire was sent to a sample of 600 women. About 16.8 per cent responded. A follow-up questionnaire drew in another 34.1 per cent. Then a probability sample of the remaining nonrespondents was contacted by telephone, and 97.2 per cent of those called co-operated.

It was found that only those interested in, and familiar with, the topic responded to the original questionnaire. Thus, it was pointed out by those who conducted the study that "if we had stopped with the original returns . . . it would have been claimed that almost twice as many women knew about the program as really do."⁶

⁶ Edward A. Suchman and Boyd McCandless, "Who Answers Questionnaires?" *Journal of Applied Psychology*, Vol. XXIV (1940), p. 760.

The problem of nonresponse—its causes and remedies—has been plaguing planners of statistical studies for many years. Much research, therefore, has been done in this field, and a vast literature exists on the subject. It often is worthwhile to consult someone familiar with this literature in the planning stage of a statistical study.

9.5 Biased Observations

In the last section we considered the possibility of missing observations—i.e., nonresponse. Now we shall turn to a brief discussion of the possibility of biased observations—i.e., incorrect responses. For example, suppose that we were considering a population of diameters of steel rods. Our discussion to this point has assumed that for any rod selected in a sample, its true diameter will be observed. Similarly, assume that we are interested in a population of ages of employees and that a sample is drawn from this population. Unless the true ages are reported, we have a sample including not true population values but rather biased observations.

It is unfortunate that biased observations are more commonplace than is generally recognized. They may result from many causes, several of which we shall survey in this section. As you shall see, some of these factors are more important than others in certain types of studies. Whatever their cause, however, it is necessary to keep in mind that biased observations present risks of inaccurate information and, hence, risks of incorrect decisions to a decision maker. Therefore, once again the problem in any statistical study is to try to minimize their occurrence by carefully planning the procedures to be used in making observations.

Consider, as examples of biased observations, the results of weighing with an “unbalanced” scale or the results of measuring with an “inexact” instrument. Such results would be biased because of an intrinsic imperfection in the method of observation. Note in particular that nonsampling errors such as these can be eliminated only by a careful testing and analysis of the method of observation. They will not tend “to even out” because of any assumed “law of averages.”

In contrast to systematic errors of measurement, it should be recognized that any method of measurement also will be affected by chance or random errors. Such errors may be due to air currents, uncontrollable variations of temperature, vibrations of the measuring

instrument, etc. They often vary in a random manner from measurement to measurement and, hence, may be analyzed from the point of view of random sampling theory.

Example 9.11 Consider a decision problem, perhaps involving the grading of a batch of a chemical compound for which observations are made by laboratory analysis. For any given laboratory procedure there will be some chance variation. Thus, if a sample of the compound were tested over and over again in essentially the same manner, different results would be obtained. Why? Because a multitude of "chance" factors would influence the results. One test might indicate a 2.31 per cent alcoholic content, a second test might show a result of 2.17 per cent, a third test a content of 2.25 per cent, etc. This chance variation may be regarded as sampling error—the result of applying the laboratory analysis only a few of the unlimited number of times it could be applied.

However, unless on the average the laboratory procedure yielded the true measurement, we would say that it also resulted in a systematic error or bias. That bias would be due not to chance but to an imperfection inherent in the method of analysis. Of course, bias, or the lack of it, can be determined only by comparing a given method of measurement with some other, supposedly unbiased, method.

Similarly, methods of measuring length, width, height, weight, volume, and so forth, may result in chance errors *and* systematic errors. Remember, then, that while random errors of measurement may be regarded as "sampling" errors, systematic errors of measurement must be regarded as biases.

To consider somewhat different types of biased observations, let us turn to statistical surveys of human populations. Three simple, commonplace, and important causes of biased observations in such studies are (a) the questionnaire, (b) interviewers, and (c) auspices. Therefore, a brief discussion and a few examples of the influence of these three factors will be instructive.

To elicit accurate information from people, a question must be framed carefully. First of all, it must be adapted for the method of making the observations in a study—that is, for a personal interview, for telephone calls, or for a request by mail. The wording of questions must be simple and unambiguous. It must not be leading. An attractive and clear form in mail questionnaires is important. Very often the preparation of a good questionnaire requires expert assistance. In most surveys in which a questionnaire is used it is desirable to pretest it. The following examples indicate some of the many possible deficiencies in questionnaires.

Example 9.12 A public opinion survey might request an answer to the question: "Are you in favor of changing the filibuster system in Congress?" While this question appears reasonable, a study once indicated that only 48 per cent of the people questioned knew the meaning of the term "filibuster."

The moral: In designing questions, use simple words which are familiar to potential respondents.

Example 9.13 In a certain study homeless men were interviewed. When asked, "Are you married?" most replied, "No." However, when the question was asked indirectly as, "Where is your wife?" a far larger proportion indicated that they were married.⁷ Such indirect questions often are good ways of getting valid information.

Example 9.14 Interviewers for the "Monthly Report on the Labor Force" originally asked: "Was this person at work on a private or government job last week?" The question was interpreted by many respondents to exclude part-time workers. The question was then changed to: "What was the person's major activity during the week?" When a person's activity was something other than working, the interviewer also inquired whether the person "did any work for pay or profit during the week?"

It was found that the new form of questioning increased the estimate of the number of working males by 900,000 and of the working females by 700,000. The additional male workers were mainly students, and the additional female workers were primarily housewives. More than half of these had worked more than 35 hours during the week in question.⁸

Thus, one always must be on guard to make sure a question actually is yielding the desired information.

Too often, it has been found, interviewers influence the observations being made. Thus, some interviewers may cause, unconsciously, human respondents to take sides with them or to take sides against them. Bias attributable to the interviewer may arise from his political or social or religious beliefs which he superimposes on respondents in the way he asks questions or in the way he treats answers. In addition, bias due to interviewers often is introduced (1) by improper training; (2) by failure of interviewers to follow instructions as to questioning, as to areas to cover, and as to treatment of nonresponse; (3) by failure of interviewers to understand the purpose of a study; or (4) by interviewer fatigue.

⁷ F. Stuart Chapin, *Field Work and Social Research* (New York: D. Appleton-Century-Crofts Co., Inc., 1920), p. 172.

⁸ Gertrude Bancroft and Emmett H. Welch, "Recent Experience with Problems of Labor Force Measurement," *Journal of the American Statistical Association*, Vol. XLI (1946), pp. 303-12.

Example 9.15 A classic and often-quoted illustration of bias arising from interviewers is provided by a 1914 social study of 2,000 destitute men. The reasons given by many of these men for their situations varied directly with the interests of their interviewers. Thus, the responses recorded by an interviewer with socialist tendencies showed a strong tendency to place the blame on industrial conditions. In contrast, the men interviewed by a prohibitionist showed a strong tendency to place the blame upon drink.⁹

To minimize interviewer bias, it is necessary to have a well-trained group of interviewers with specific instructions to follow. This ordinarily is possible, for most problems, only in sample studies which require only a small corps of interviewers. Large-scale surveys which require a large staff of interviewers increase tremendously the difficulty of training and offer more opportunity for variation in the quality of interviewing. This is one reason why, in Chapter 4, it was pointed out that samples may result in more accurate information than complete counts.

At times it has been found that people will report differently, if at all, because their attitudes are influenced by the organization conducting the study, or by the purpose of the study. This is called the bias arising from auspices. Thus, it has been found that some persons will report differently, either unintentionally or intentionally, to studies carried out by governmental agencies and to studies carried out by private research organizations.

Example 9.16 A number of years ago a census in a Chinese province, for military conscription and taxation purposes, found a population of only 28 million. Another survey in the province, a few years later, but for the purpose of granting aid for famine relief, found a population of 105 million.

Example 9.17 One statistician once related the following incident: "I remember a case in California not long ago, where they were talking about the closing of banks on Saturday. Two surveys had been run. One survey had been run to ask the general public about the employment policies of a certain bank. In that survey, they asked whether the bank should be closed on Saturdays. Over-whelmingly, the people said, 'Sure, yes, they should be closed on Saturdays.' Apparently the reason was that it would be better for the employee, who would get the day off. The other survey talked about services to the public; and whether banks stayed open the right hours. They asked the same question, 'Should the banks stay open on Saturday?' Just

⁹ Stuart A. Rice, "Contagious Bias in the Interview: A Methodological Note," *American Journal of Sociology*, Vol. XXXV (1929), pp. 420-23.

the opposite results came in. The public said the banks should stay open on Saturday to give better service to the public."¹⁰

The results were inconsistent for the respondents had different frames of reference.

Even if all the afore-mentioned causes of biased observations are minimized, inaccurate responses will still occur for other reasons. Some responses reflect the lack of ability of people to provide accurate information; some others, the refusal to do so. Thus, biased observations may result from memory failures, guessing, accidental mistakes in responding, and "prestige bias." The latter is the term used to explain wrong answers arising from a respondent's pride—his reluctance to admit he reads a tabloid newspaper, or his downgrading of his age, or his upgrading of his salary.

Example 9.18 One recent survey of readership and recognition of financial and business writers for New York newspapers indicated the following response bias. People were asked, among other things, if they read the column written by Kenneth Albridge. Of the respondents, 5.7 per cent claimed to have read it at least occasionally. Several named the newspaper, a feat all the more remarkable because Kenneth Albridge was a fictitious columnist. Undoubtedly, some of this bias resulted from an accidental mistake in responding; while others resulted from an inability to say "I don't know." The latter reflects a prestige bias. Incidentally, the purpose of that particular question was to provide some information on response bias.

Example 9.19 A study on the ability of British government employees to recall their absences during a year showed the importance of the memory factor in statistical surveys. Some employees overstated the amount of sick leave taken. Others underestimated it. Many others, although they could recall being out, were unable to remember when they were absent.¹¹

Example 9.20 In the 1950 *Survey of Consumer Finances* homeowners were asked to estimate the market value of their houses. Later, estimates of the same houses were made by professional appraisers. The discrepancies between the two sets of estimates were large. Thus, 63 per cent of the homeowners' estimates differed by more

¹⁰ J. Stevens Stock, "Non-sampling Errors in Sampling for Business," *Proceedings of the Conference on Business Application of Statistical Sampling Methods*, (Monticello, Ill.: The Bureau of Business Management, University of Illinois, 1950), p. 120.

¹¹ Percy G. Gray, "The Memory Factor in Social Surveys," *Journal of the American Statistical Association*, Vol. L (1955), pp. 344-63.

than 10 per cent from the corresponding professional appraisals. However, on the whole, these individual biases tended to cancel out.¹²

Careful planning always helps to avoid biased observations. Yet some bias may still creep into any study. On the basis of their past experiences, statisticians and subject-matter experts often may be able to estimate roughly the potential magnitude of these biases in a given study. If they will be excessive and cannot be controlled, the only solution to the situation may be to give up the study—to reformulate the problem so that another, more accurate decision procedure can be used to solve it.

9.6 Biases in Data Flow

In the next chapter we shall study the details of data flow. This topic deals with the collection of observations, the processing of these data, their tabulation and summarization, and any preceding or subsequent arithmetical calculations. This section anticipates some of the difficulties—some of the potential opportunities for biases—in data flow.

Thus, one source of this type of trouble in many surveys is a disorganized collection procedure. Poor supervision of interviewers presents an illustration. Such a difficulty may result in ill-trained interviewers and, hence, interviewer bias. Or it may result in observations being collected twice, or not at all, from an elementary unit that is drawn into a sample. Poor record keeping or inadequate forms for the collection of observations also provides an opportunity for bias in data flow. For an illustration of the consequences of a disorganized collection procedure, Example 9.9 may be reread at this point.

In some studies it is necessary to edit and to code observations after they are collected but before they are tabulated and the pertinent statistics computed. For example, suppose that one marketing survey includes the question: "What features of this product appeal to you?" Before answers are tabulated, they must be coded—fitted into various groupings. However, two coders or a coder and his supervisor may disagree on the particular grouping to which a given answer should be assigned. And, in any event, the opportunities for bias from inaccurate coding are evident. Where coding is an important step in data flow, precise instructions on how to code

¹² Leslie Kish and John Lansing, "Response Errors in Estimating the Value of Homes," *Journal of the American Statistical Association*, Vol. XLIX (1954), pp. 520-38.

must be given workers. This also is another situation where sampling offers an advantage over a complete count—a smaller, better trained work force to do more accurate work is possible with a sample.

Arithmetical mistakes in tabulation and in calculation also represent potential sources of bias in data flow. Such mistakes usually should be negligible, but at times they are not. However, they ordinarily may be minimized by control totals and other numerical checks. Then, too, it usually is possible and desirable to use a sampling plan to control clerical accuracy.

In conclusion, it should be noted that the insidious thing about any and all biases is their constancy and the consequent difficulty of detecting them. Furthermore, biases have a tendency to cumulate rather than to compensate one another. Our examples in this chapter point up situations where biases were found to exist or where they may be expected to exist. In how many other situations are they not detected—nor any thought given to the problem of detecting them?

9.7 You Should Now Know That

To check the trustworthiness of the results of a statistical study, ordinarily one must check the procedure by which the data were collected.

Procedural biases, or nonsampling errors, are imperfections inherent in the procedure used to elicit information.

Biases will arise even though a complete count is taken.

Biases usually are impossible to measure and often difficult to detect.

In contrast, sampling error is the difference between the result from a sample and the result which would have been obtained from a complete count taken with the same care as the sample. The risks of (probability) sampling errors are measurable.

Almost every study is subject to some sampling error, some procedural bias, or, more usually, both.

Generally, the only way of controlling biases is by planning a study carefully to limit the opportunities for their occurrence.

Knowledge of ways in which biases may occur is therefore important.

Biases may be due to the improper formulation of the decision problem—defining the population poorly, specifying the wrong decision-parameter, securing an inadequate frame, and so forth.

Biases may arise because of an imperfect sampling procedure or a

failure to understand correctly the risks involved in applying a sampling procedure.

Bias may result from missing observations (nonresponse) or from biased observations (inaccurate responses).

The bias of nonresponse cannot be alleviated by increasing the sample size.

Biased observations may result from a poorly designed questionnaire, an ill-trained interviewer, the auspices of the study, a respondent's memory failure, prestige bias, and so forth.

Biases in data flow may result from an unorganized collection procedure, faulty editing, or coding of responses and arithmetical errors.

9.8 You Should Now Be Able to Solve

1. Discuss the following statement from the point of view of the subject matter of this chapter:

A statistical inspection procedure, no matter how perfect it is statistically, is no good without the proper supervision.

2. In a study of the burden of a gross receipts tax upon the firms that pay the tax, the following question was asked: "Do you pass the cost of the tax on to purchasers of your products?" In what ways could this question lead to biased responses? Explain.

3. A study of retail trade asks respondents to report sales in thousands of dollars. How might this request be a source of bias?

4. It has been determined that a sample of 600 firms is needed to estimate the potential market for a new lubricant. A mail questionnaire is to be used. Past experience has indicated that a non-response of about 25 per cent is to be expected. Therefore, a random sample of 800 firms is selected to assure a sample of 600 respondents. Evaluate this procedure.

5. A mail questionnaire was sent to a random sample of 4,000 individuals to determine their attitudes toward a new type of popular science magazine. Questionnaires were tabulated as they were received. The results for the first 600 returns were as follows:

<i>Questionnaire Checked</i>	<i>Proportion Favoring the Magazine</i>
200.....	.50
400.....	.54
600.....	.52

Since the proportion remained fairly stable, it was decided not to tabulate the remaining questionnaires. Thus, additional processing expenses were saved.

Evaluate this procedure.

6. The table below shows the relative frequency for digits of age reported in the Census of Population for 1880 and 1950:¹³

Digit of Age	1880	1950
0.....	16.8	11.2
1.....	6.7	8.9
2.....	9.4	10.2
3.....	8.6	9.7
4.....	8.8	9.7
5.....	13.4	10.6
6.....	9.4	9.8
7.....	8.5	9.7
8.....	10.2	10.2
9.....	8.2	10.1
Total.....	100.0	100.0

Analyze this table for possible indications of bias. Discuss your findings.

How might these biases be overcome?

7. A simple random sample of 300 adults is to be selected in a given town. An available list of all the families in the town is to be used as the frame. It is decided to select a simple random sample of 300 families and to interview one adult in each family.

Is this sampling procedure biased? Explain.

8. You are to conduct a poll to determine voting intentions in the next election. A random sample of households is to be visited, and husbands and wives are to be interviewed in each household. In the interest of speed and economy, husbands and wives are to be interviewed together.

Evaluate the suggested procedure.

9. An interviewer is to visit the homes of a randomly selected sample of families. If there is no one at home, he is to visit a family next door. However, to avoid bias, he is to alternate his visits between the neighbor to the right and the neighbor to the left of the not-at-homes.

Evaluate this sampling procedure.

9.9 You Will Also Find That

Discussion, analysis, and examples of biases are given in:

DEMING, W. EDWARDS. *Some Theory of Sampling*, pp. 24-52. New York: John Wiley & Sons, Inc., 1950.

¹³ Robert J. Myers "Accuracy of Age Reporting in the 1950 United States Census," *Journal of the American Statistical Association*, Vol. XLIX (1954), pp. 826-31.

HANSEN, MORRIS H., HURWITZ, WILLIAM N., and MADOW, WILLIAM G. *Sample Survey Methods and Theory*, Vol. I, pp. 56-92. New York: John Wiley & Sons, Inc., 1953.

PARTEN, MILDRED. *Surveys, Polls and Samples: Practical Procedures*, pp. 403-24. New York: Harper & Bros., 1950.

An interesting article on bias resulting from incorrectly formulating a problem is:

KIMBALL, A. W. "Errors of the Third Kind in Statistical Consulting," *Journal of the American Statistical Association*, Vol. LII (1957), pp. 133-42.

For further description of misuses of statistics see:

HUFF, DARRELL. *How to Lie with Statistics*. New York: W. W. Norton & Co., Inc., 1954.

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics: A New Approach*, pp. 64-89. Glencoe, Ill.: The Free Press, 1956.

Data Flow

10.1 Data Flow Defined

In the preceding chapters we have considered the steps necessary in the planning of a statistical study. These steps may be summarized as follows:

1. Identify the problem in terms of alternative courses of action, states of the world, and possible consequences.
2. Define the problem in statistical terms; i.e., define the population of study and specify the pertinent decision-parameter(s).
3. Determine whether a decision can be reached without the benefit of observations or on the basis of available secondary data. If not, determine whether the problem requires a complete count, a judgment sample, a convenience sample, or a probability sample.
4. If a probability sample is to be used:
 - (a) Set the risks that you are willing to take in relying on sample information as a basis for action.
 - (b) Determine the decision rule that will keep the risks of incorrect action within desired limits.
 - (c) Plan the decision procedure so that procedural biases (nonsampling errors), as well as sampling errors, will be controlled.

Thus, we have discussed how one determines what type of information is wanted, how it will be used, and, moreover, how one may plan a statistical study so that when it is carried out, the results will serve as a guide to a decision.

In this chapter we shall survey several of the major steps and problems in the implementation of a statistical study. Once elementary units are drawn for a specified purpose, data on their characteristics of interest must be collected, these data processed and tabu-

lated, the pertinent statistics calculated, the appropriate action taken, and a final report on the study written. This series of steps may be called *data flow*. It deals with the flow of information from the time that the sample of elementary units is chosen to the time that the observations are presented in final form to serve as guides for decision.

Unless the flow of data is efficient and effective, the cost and the time necessary to conduct a study are enormously increased. Furthermore, as we know from Chapter 9, data flow, unless well planned, can be the source of several procedural biases which only add unnecessarily to the risks of relying on the results of a study.

10.2 Methods of Data Collection

Because we are concerned with the flow of data, it is appropriate to begin our studies with some discussion of methods of obtaining pertinent information from the elementary units under investigation.

We may first note that in some studies, the method of data collection is simple and straightforward—observations can be made easily by direct observation. For example, to determine whether or not a glass jar is broken, we ordinarily need only to look at it; while to verify the diameter of a bearing, we need only to measure it.

In other problems, the method of data collection may be a far more complex procedure. For example, it may require a lengthy laboratory analysis or an involved physical test of materials. It may necessitate the use of complex equipment—for example, X-ray or fluoroscope inspection often is used to determine whether or not flaws exist in aluminum castings.

As our introductory comments indicate, the exact procedure to be used to collect data from elementary units in any study depends on the nature of the observations, and, moreover, on the nature of the problem. Furthermore, it must be pointed out that the choice of a method is almost exclusively the responsibility of an expert in the subject matter; e.g., the engineer, the chemist, or the psychologist. It is not, obviously, a statistical problem, per se. Yet it is an important problem and must be given serious consideration in almost every statistical study.

A discussion of most methods of data collection would go beyond the scope of this textbook. However, in the remaining paragraphs of this section we shall consider various methods of data collection which are of importance in statistical studies dealing with human

populations and, hence, which arise very frequently. The possibility of inaccurate observations in such studies is great, as we have seen in Chapter 9. What, then, are the ways of collecting observations from people so that the risks of biased responses are minimized?

Among the methods of data collection often used in surveys of human populations, the following are most important: (1) personal interview; (2) mail questionnaire; (3) telephone interview; (4) panel technique; and (5) recording devices. There is a vast literature on the use of each of these methods. Some of the literature is listed in the last section of this chapter. Here we can only discuss some of the characteristics of each method.

There is no one best method for obtaining information. The method used will depend on such factors as the nature of the problem, the population, the time limit, and the available funds. Very often two of the methods can be combined. In any event, great care must be used in eliciting information by means of these methods, for improper techniques give rise to biases and, hence, affect the accuracy of a study.

1. Personal Interview. In the personal interview, questions are prepared in advance and are printed on a schedule on which the answers also are recorded. This method results in a relatively high percentage of usable returns from persons approached. Personal contact enables the interviewer to clarify any misunderstanding of the questions. Sometimes he may be able to make observations to check on the accuracy of answers. Thus, a housewife might claim preference for one brand of coffee while another brand is visibly in use. Sensitive questions can be interspersed strategically, and the respondent's time can be utilized more effectively. By allowing more time for thought, recall can be facilitated.

The feature of personal contact, which makes the personal interview so desirable, is also a cause of its disadvantages. An interviewer must be careful not to allow his personal judgments to influence the respondent. The inflection of his voice in reading a question, a move of an eyebrow, a smile, or changing the wording of the question might exert an influence. Careful training and supervision of interviewers, therefore, is important. This requires a complex organization. Such factors tend to make the personal interview a costly method. However, in considering cost, we must remember that the amount of nonresponse is much less than under other methods. Hence, fewer call-backs will be necessary. Thus, when we consider costs in terms of usable information obtained, the costs of

personal interviews might not loom so large in comparison with other methods.

2. Mail Questionnaire. The mail questionnaire is the most widely used method of data collection in survey work. It consists of a series of questions sent to the respondent and returned by him by mail. A mail questionnaire is of relatively low cost when considered in terms of questionnaires sent out. No staff training of interviewers is necessary, and a standardized set of questions can be prepared. Mail questionnaires avoid the possibility of interviewer influence upon the respondent. In the privacy of his own home, frank answers can be written at the respondent's own convenience. A shortcoming of the mail questionnaire is the high proportion of nonresponse. You are aware of this because no doubt you too often have filed such questionnaires in the waste basket. This weakness may make the costs of mail questionnaires relatively high when considered in terms of usable completed questionnaires. Returns of questionnaires from the general public range from 10 to 20 per cent. Returns to special groups or to governmental agencies range higher. Thus, the 1954 *Census of Manufactures* was conducted through a mailed questionnaire, with more than a 90 per cent return. However, no matter what the rate of return, the failure to investigate nonresponse may result in a serious bias in the results. A mail questionnaire can be supplemented by follow-up letters, and then by interviews, to obtain answers from a properly drawn sample of nonrespondents. The combination of mail questionnaire and personal interview is about the best and most economical combination. It results in relatively low costs per unit of usable data obtained if the follow-up does not have to be extensive. At the same time the problem of nonresponse is alleviated through the personal interviews.

Example 10.1 The use of a combined mail questionnaire and interview was reported recently in *Applied Statistics*, a British statistical publication. The Ministry of Agriculture and Fisheries desired to determine the number of poultry and pigs kept in gardens and other nonagricultural holdings. No statistics had been gathered since feed rationing had been removed in August, 1953. The Ministry wanted to determine the impact of the discontinuance of rationing on animals raised on nonagricultural holdings to see whether action to control the growth of such holdings was necessary. A letter accompanied the questionnaire and was followed up by two reminder letters. The nonrespondents were then interviewed. In rural areas the response by mail was 87.9 per cent of the sample. Supplementary in-

interviewing raised this total to 99.5 per cent. In urban areas the corresponding figures were 94.1 per cent and 99.8 per cent.¹

3. Telephone Interview. Costs of telephone interviews are relatively low, and a wide area can be covered without adding greatly to the expense. It is easy to train and supervise interviewers and to develop a standardized approach. Furthermore, it has been found that the refusal rate is low—usually not more than 2 to 3 per cent. Its main applicability is to radio and television audience studies. A person is asked the name of the program that he is listening to and usually some questions about the product that is being advertised.

Example 10.2 The method of telephone interviews for measuring radio (and also television) audience appeal was developed largely by C. E. Hooper, Inc.² The “Hooperating” was used as a guide in making decisions by both buyers and sellers of air time. The questions asked of radio listeners were brief and only five in number.

1. Were you listening to the radio just now?
2. To what program were you listening, please?
3. Over what station is that program coming in?
4. What advertiser puts on that program?
5. Please tell me how many men, women, and children were listening to the radio when the telephone rang?

The telephone interview can be effective only when the questions are few and not many details are required. It is difficult to detect misinformation. Most serious of all, however, is the fact that even though refusals are few, there are many no-answers, busy signals, and wrong numbers. At times these amount to one-third of the calls made. This creates a large amount of nonresponse, especially if no call-backs are made. The use of telephone interviews also limits the frame since nontelephone owners are excluded from being chosen for the sample. Care must, therefore, be taken to ascertain whether telephone owners provide an adequate population for the problem in question.

4. Panel Technique. This method involves the selection of a group of respondents who are to be surveyed from time to time. Thus, we might have a consumer panel consisting of a few thousand women who are asked to report periodically on their reactions to forms of advertising, or to preferences for new products. The data are used to guide decisions as to advertising media and product de-

¹ Percy G. Gray, “A Sample Survey with Both a Postal and an Interview Stage,” *Applied Statistics*, Vol. VI (1957), pp. 139-53.

² Matthew N. Chappell and C. E. Hooper, *Radio Audience Measurement* (New York: Stephen Daye Press, 1944), p. 62.

sign. The method also is used by the federal government to determine consumer behavior patterns. In this latter instance panelists are required to keep records of their purchases. The information derived in this manner is used to determine the relative importance of the various components of the Consumer Price Index (food, rent, furniture, clothing, etc.).

The use of panels facilitates the study of changing attitudes of consumers. If different samples are chosen at different times, it is difficult to assess whether any changes reflect a shift in attitude or result from the use of different samples. The usefulness of a panel requires that it be kept intact. This has been one of the great difficulties with the panel technique. Panelists change their residences, keep faulty records, change their minds in midstream, or just get "bored with the whole idea." Another serious shortcoming of the panel technique is the psychological effect that being a participant has upon many panelists. It has been claimed that panelists frequently develop a critical "set"—or attitude—to questions. Panelists, having stated an opinion, often try to stick to it. They may thus be less subject to changes of opinion than the general public as a whole.

5. Automatic Recording Devices. In keeping with an era of automation, various devices have been perfected to record observations automatically. The photoelectric cell has been used to record the number of cars using municipal parking lot facilities at various times during the day. Such information is useful in making decisions on fees and maximum allowable parking time. It also can be used to determine shopping patterns in stores during the course of the day. Other interesting devices have been perfected for radio and television audience surveys.

Example 10.3 The A. C. Nielson Company utilizes a device called an "Audimeter" that is attached to a set. The Audimeter records once every minute to show whether the set is on or off, as well as the station to which the set is tuned when it is on. The Audimeter record tape is reviewed every month.

Example 10.4 The Lazarsfeld-Stanton program analyzer is a device for determining what features of different programs appeal to radio and television audiences. The listener records his reaction by pushing a green button to signify "like," a red button to signify "dislike," and by not pushing either button to register "indifference." A permanent record of these reactions is kept. An interview usually accompanies the use of the analyzer to determine the reasons for the recorded reactions.

The advantages of these contrivances are that they do not depend upon memory nor are responses influenced by interviewers. Data can be obtained over an extended period of time without imposing upon the respondent. On the other hand, these devices are rather costly, and the size of the sample is limited. It is difficult to obtain a probability sample of listeners, as all families selected by a random process might not be willing to permit the installation of the devices. Merely to record the opinions of families that do permit installation may result in biased information. Furthermore, the devices merely record that the set was tuned in on a given station. It does not follow that people are listening when a set is in use.

10.3 Organization of the Collection Procedure

Whether observations are to be collected by direct observation, by personal interviews, by mail questionnaires, by telephone queries, by some mechanical device, or by some other method, the collection procedure must be well organized. Of course, the exact nature of the collection procedure must be tied to the method of data collection to be utilized and to the special problem at hand. In this section we shall indicate only some of the more general requirements of an efficient system for collecting observations.

First of all, it must be noted that the proper records and forms are of importance. Thus, in an industrial inspection problem, a written record of the number of defectives and the number of nondefectives, or of the measurements made, must be maintained for any acceptance sampling plan to be effective. To rely on an inspector's memory in place of a written record would be foolhardy indeed. Similarly, in an inventory count, the proper forms will facilitate the collection of accurate observations.

For statistical studies of human populations, for which the personal or telephone interview method of data collection is used, there must be a suitable schedule. For other studies, such as those in which persons are contacted by mail, there must be a suitable questionnaire. (Strictly speaking, a questionnaire is a form distributed through the mails or given to respondents to be filled out without any assistance; a schedule is a form filled out by an interviewer or, at least, in his presence. Often the terms are used interchangeably.) The collection of data and subsequent processing of these data can be facilitated by the proper schedule or questionnaire. Therefore, several considerations in their design must be borne in mind:

1. The physical appearance of the schedule or questionnaire may affect the amount and the accuracy of response.
2. Questions should be as concise as is possible; moreover, the entire schedule or questionnaire should be as concise as possible.
3. Questions must be worded clearly; they must be appropriate for the type of respondents expected to participate.
4. Only questions which bear on pertinent information should be included; they should be formulated so as to yield this information.
5. Usually questions should not be included if their answers can be obtained more accurately and more easily from other sources.
6. Always pretest the schedule or the questionnaire.

As the preceding remarks indicate, the choice of the proper forms and their design is important. These matters are often problems for specialists; almost always, they are problems for experts in subject matter rather than for the statistician.

Another important factor in the organization of a collection procedure is the hiring, training, and supervision of personnel—e.g., interviewers or inspectors. A well-organized collection procedure must be so designed that work loads are evenly distributed and reasonably scheduled. It also should include provision for (sample) checks on the actual performance of personnel.

Even though the population and the purpose of a study are most carefully delineated, the questionnaire in perfect form, and the sample design most ingenious, the accuracy of a study will be greatly impaired unless the collection procedure is well organized.

Example 10.5 Consider the experience with the census of industrial and commercial enterprises conducted in France in March, 1946. A cursory examination of the census questionnaires seemed to indicate that many firms had not answered the questionnaires. A sample survey was, therefore, conducted in urban areas of France to discover the degree of under-reporting that had taken place. From the sample study it was estimated that about 25 per cent of commercial and industrial enterprises were overlooked. Furthermore, it was found that firms of all sizes were missed. It was concluded that the high proportion of enterprises overlooked in the census resulted from a disorganized collection procedure. The result: The French government never published the results of the 1946 census of business.

This illustration emphasizes a fact that we have mentioned earlier. Since a sample study is conducted on a smaller scale than a census, it is often easier to secure greater accuracy from sample studies. This is in part attributable to the fact that a collection procedure can be better organized and controlled more effectively in sample studies.

We cannot dwell at length on all of the many aspects of the organization of collection procedures since many technical details are involved. You should be able to appreciate the complexity of this problem for many statistical studies by seeing what steps and precautions a research agency actually took in its conduct of a nationwide survey.

Example 10.6 A few years ago a statistical study was conducted by Alfred Politz Research, Inc., for *Life* magazine to measure people's interests in movies, magazines, radio, and television. The report on the survey is entitled: *A Study of Four Media—Their Accumulative and Repeat Audiences*. The design of the study began in the fall of 1951 and took more than 1 year to complete. Actual interviews began in February, 1952, and it took $2\frac{1}{2}$ years to complete the entire survey. This period of time was required because the scope of the study necessitated a very carefully planned collection procedure. Among the steps taken to achieve this result were the following:

After the questionnaire had been prepared, 1,200 interviews were conducted in a pilot study in the New York metropolitan area. The pilot study had as its aim to test each factor which might affect the accuracy of the study.

After the pilot study, a training program for interviewers began. Thirty-five instructors were used to train interviewers. Each interviewer was trained personally at training centers set up throughout the United States. The trainees had the purpose of the survey carefully explained to them and were briefed on the subject matter of the questionnaire. Training in the art of asking questions also was provided. Instructors demonstrated the entire interview. Trainees went through the questionnaire with new respondents under observations of the instructor, and appropriate suggestions and criticisms were offered by the instructor.

As a final step in the training program, an actual survey situation was simulated. Test assignments were sent out to all interviewers. The interviewers went to the designated areas (none of which were used in the final survey) to locate households, to select specific individuals, and, when necessary, to make call-backs just as in an actual study. The field test was conducted under the direction of the home office which was to co-ordinate the actual survey. The completed questionnaires were sent to the home office where they were checked and critical comments made. Copies of the comments were sent to the interviewer and his instructor for discussion.

In all, 207 interviewers were employed. During the conduct of the survey checks were made on interviewers by telephone, personal visits, post cards, and letters. Over-all, 7,141 respondents were interviewed, 36,868 interviews made, and 99,052 visits required.

Obviously, most sample studies are not of the same magnitude as the one in the previous illustration. This does not mean, however,

that careful planning of the collection procedure is any less important in small-scale studies.

10.4 Processing the Data

After observations have been completed, they must be processed. Processing includes the *editing* of observations for completeness, consistency, and accuracy. After the data are edited, they are ready for *coding* or classification so that the results can be tabulated or summarized.

We say that a return is incomplete if, for a given respondent, some pertinent observations are missing. An attempt should be made to complete such returns. Failure to obtain such observations may result in a serious bias of nonresponse.

Observations also must be edited to determine their quality—to detect, where possible, biased observations. In this connection, there are several rules to follow. First of all, evidence of the quality of observations often is indicated in the data themselves.

Example 10.7 In a survey of family expenditures, a family reports the ownership of a home. However, no expenditure is reported for property taxes. This seems inconsistent, and the matter should be investigated.

This example illustrates a useful idea for many statistical studies. It often is desirable to make observations which verify other observations. Such verifications will not necessarily provide conclusive evidence of poor quality. However, if inconsistencies appear, they should arouse suspicion, and, moreover, the observations should be rechecked.

Example 10.8 A trade association conducts an annual survey of financial and operating ratios for its industry. Among the ratios reported by member firms are the following: net sales/inventory, inventory/net working capital, and net sales/net working capital. You will observe that the product of the first two ratios is equal to the third. In this manner the consistency of the report can be checked for these items. Which ratio or ratios are in error cannot be determined by the check, but questionable reports can be returned for correction.

In addition to inconsistencies, a return may contain values that are unlikely. A skilled editor can often recognize the causes of such unlikely values and take some action to correct them.

Example 10.9 In a study of financial ratios the per cent of net worth that is invested in plant and equipment is requested. In a given

instance the percentage might be reported as .208. This is a very unlikely figure for the industry. What may have happened is that the respondent forgot to multiply the ratio by 100 to convert the answer to a percentage.

At times discovery of the cause of an extreme value is not as simple as in the previous illustration. In such instances, a decision must be reached in the face of uncertainty as to whether to retain or discard such extreme items.

Example 10.10 A review of the results of a time study on an assembling operation showed the following times (in minutes):

1.14	1.07	1.21	1.18	1.60
1.10	1.18	1.02	1.11	1.25

The time-and-motion-study man, in editing these data, *may* be suspicious of the recorded time of 1.60. A check reveals no clerical errors or apparent unusual circumstances during the time of the operation. However, this number still *may* be the result of an out-of-control condition, not a normal condition. As such it would not be part of the sample from the population of interest. This population is the set of times which would occur if the operation were repeated over and over again under stable conditions.

One solution to the problem is simply to discard the suspected observation. This is a good solution *only if* two rules are followed. First of all, the decision to discard any observation should not be made after the data are collected—it should be made a priori by the formulation of a statistical decision rule. Otherwise, the sample data may influence not only the decision itself but the rule for making the decision, and the sample study may wind up showing what someone wanted it to show. If so, there is no point in doing the study in the first place. Secondly, such an a priori decision rule must be objective—that is, it must be based on a knowledge of the risks involved in the decision on whether or not to discard the extreme time. Thus, two errors may be recognized, namely:

1. Discarding an observation when it should not be discarded because it is not biased and merely large or small because of chance.
2. Not discarding an observation when it should be discarded because it is biased and is not extreme due to chance factors.

The situation thus presents a testing problem and may be treated by methods similar to those discussed in Chapter 8.

Unusual regularity in the responses may sometimes be observed in editing. This may indicate inaccuracies, and the returns should be reviewed.

Example 10.11 In the 1900 *Census of Population* it was found that the number of Negroes listed as unable to speak English was suspiciously large. The matter was investigated, and it was found

that the schedule, in contiguous columns, contained questions on the ability to read, ability to write, and ability to speak English. In the case of white respondents the answers usually were "yes, yes, yes."

The pattern of recording the first response three times was carried over to interviews with Negroes who could neither read nor write English. Often after eliciting the "no" response to the first question on ability to read, careless interviewers would write "no, no, no" without asking the second and third questions. In many instances the replies would have been "no, no, yes."

While the "yes, yes, yes" was a regularity in the recording of responses that would not arouse suspicion, the recording of "no, no, no" regularly indicated a possible inaccuracy. The tabulations for Negroes, therefore, had to be abandoned for that census.³

The afore-mentioned checks for accuracy of data are applicable not only in editing but should be used to appraise numerical data wherever you may happen to encounter them.

Example 10.12 A statistician made a survey⁴ of the numerical data cited in the Kinsey report on males. On page 10 of the report it is stated that the number of individuals involved in the study was 12,214. On page 5, however, a map of the United States shows the parts of the country from which the respondents came. The map contains 427 dots, and it is stated that each dot represents 50 people. This indicates that the study included 21,350 individuals rather than 12,214.

After observations are edited, it may be necessary to code them. Coding is the classification of observations to facilitate their tabulation or summarization. The process of coding is particularly important for statistical studies in which there is the possibility of many different observations. These many different observations must be classified in order to have meaning.

Thus, consider the use of a questionnaire. Questions generally can be divided into two broad groups: (1) fixed-alternative questions, and (2) free-response or open-ended questions.

Example 10.13 Consider some questions that might appear in a marketing research study of consumers' smoking habits. An illustration of a fixed-alternative question is:

Do you have a favorite brand of cigarettes?

1. Yes ☐

2. No ☐

³ Cited in Mildred Parten, *Surveys, Polls and Samples: Practical Procedures* (New York: Harper & Bros., 1950), p. 432.

⁴ W. Allen Wallis, "Statistics of the Kinsey Report," *Journal of the American Statistical Association*, Vol. XLIV (1949), pp. 463-84.

There may also be more than two alternatives as illustrated below:

How many packs of cigarettes do you smoke in a day?
Less than $\frac{1}{2}$ — $\frac{1}{2}$ — 1— $1\frac{1}{2}$ — 2 or more—

Let's look at an open-ended question:

What features of this cigarette advertisement appeal to you?

There are no fixed choices of answers to this question. The respondent answers and the interviewer records the reason.

Some questions combine both methods.

Do you favor a flip-top cigarette box?

1. Yes
2. No
3. No preference

If your answer is "yes" or "no," state your reasons.

The first part of the question offers fixed alternatives, while the request for reasons illustrates an open-ended question.

The coding of fixed alternative questions is not always necessary if hand tabulation is to be used. The answers on the questionnaire may be used as a basis for sorting the responses. Thus, papers with "no" responses may be put on one pile and "yes" responses on a second pile. When replies are to be tabulated by machine, numbers or letters must be assigned to the different alternatives. The code numbers may appear on the questionnaire as in the illustrations above. Thus, in the question concerning preferences for a flip-top box, a 1 would represent "yes"; a 2, "no"; and a 3, "no preference." At times the code numbers can be checked off on a specially prepared coding card. Such coding cards also could be used for hand tabulation since it is much easier to sort cards than questionnaires.

The coding of open-ended questions is a more complex operation. Categories must be determined so that each answer fits only one category and so that a category exists for each response. In coding open-ended questions it is customary to include a miscellaneous category. However, this category must not be too inclusive. Thus, a report from one survey on sources of local government revenue classified approximately 36 per cent of all tax receipts as miscellaneous. Greater care in the design of appropriate categories would have reduced the magnitude of the miscellaneous group and would have offered additional pertinent information. Obviously, in all

cases very precise instructions and training must be provided for coders.

10.5 Tabulation of Data

The tabulation of data involves the summarization of results in statistical tables. These tables may be employed for the purpose of computing statistics required by a decision rule, or for presenting numerical information in a final report.

Plans for the tabulation of observations should be made at the time the questionnaire is designed. In fact, the design must take into consideration the ease of tabulating the responses. Such planning includes determination of the information to be contained in a table, its layout, and detailed instructions for tabulators.

A form of table commonly used to summarize numerical data is the frequency distribution. We already have discussed absolute and relative frequency distributions. We now shall consider some methods for the construction of such distributions. The tallying of the occurrences of various attributes or of variates can be performed either by hand or by machine. The greater the amount of data to be processed, the greater is the advantage of machine tabulation. Machine methods will be considered in a later section. For the present, we illustrate an efficient procedure for the tallying of frequencies by hand.

Example 10.14 The following 15 responses were received to the request for monthly rentals (in dollars) of dwelling units: 67.50, 75, 74.60, 78, 69, 68.50, 73, 72, 77, 73, 74, 78, 67, 72, 71.50. The technique of cross-fives is a convenient method to count frequencies. Four vertical strokes represent four frequencies and a diagonal line crossing them is the fifth, e.g.:

Average Monthly Rent	Number of Dwelling Units
\$65-\$69.99.....	//// 4
\$70-\$74.99.....	/ 7
\$75-\$79.99.....	//// 4
Total.....	15

The construction of a frequency distribution involves the determination of the best method of grouping the data into classes. Thus, in Example 10.14, monthly rentals are divided into three classes. An important factor in the process of grouping data is the choice of *class limits*—the lowest and highest values of observations that may fall in any one class. Thus, in Example 10.14, the limits of the first class are \$65.00 and \$69.99. Class limits must be chosen so that they are mutually exclusive. That is, each item should fit into one

and only one class. Thus, the class limits shown below are *not* mutually exclusive, and should never be used.

\$1,500–\$2,000
\$2,000–\$2,500

We could not determine whether a \$2,000 item belongs in the first or second class. Another consideration is the *class interval*—the range in values for each class. The class interval is equal to the difference between two consecutive upper limits (or between two consecutive lower limits). Thus, in Example 10.14 the class interval is \$5.00—\$70 minus \$65 or \$74.99 minus \$69.99.

The choice of the classes, the class interval, and the class limits will depend on the purposes that the frequency distribution is to serve.

Example 10.15 A buyer in a department store desires to know how many men's dress shirts of each size he is to keep in stock. A relative frequency distribution of men's shirt sizes would enable him to make a decision.

Size	Relative Frequency
14	
14½	
15	
15½	

In this situation, the classes represent a single size. It would not be helpful to group the sizes.

By contrast, consider a buyer who must determine the number of sport shirts of each size to keep in stock. Sport shirts come in the following sizes: small (14 and 14½), medium (15 and 15½), large (16 and 16½), and extra-large (17 and 17½). It therefore would be desirable to group the sizes as follows:

Size	Relative Frequency
14–14½	
15–15½	
16–16½	
17–17½	

While the basic criterion in setting up a frequency distribution is the purpose it will serve, it often is helpful to keep a few general suggestions in mind. To provide a basis for discussion of these suggestions, Example 10.16 presents a formal frequency distribution.

Example 10.16 The following table presents some of the results of the 1953 and 1958 Surveys of Consumer Finances. These surveys are conducted annually under the auspices of the Federal Reserve Board.

**PRICES PAID BY PURCHASERS OF NEW AUTOMOBILES
IN THE UNITED STATES: 1952 AND 1957**

Price (in Dollars)	Percentage of Purchasers*	
	1957	1952
Under 1,500.....	(†)	1
1,500-1,999.....	5	5
2,000-2,499.....	15	40
2,500-2,999.....	25	26
3,000-3,499.....	24	18
3,500 and over.....	31	8
Total.....	100‡	98‡

* Sample of 250 spending units for 1957 and 245 spending units for 1952.

† No cases reported or less than one half of 1 per cent.

‡ Does not add to 100 since the nonresponse was 2 per cent in 1952. Nonresponse in 1957 was 1 per cent. Data add to 100 because of rounding.

Source: Based on Supplementary Table 3 in "1956 Survey of Consumer Finances—Durable Goods and Housing," *Federal Reserve Bulletin*, August, 1956, p. 817; and Supplementary Table 6 in "1958 Survey of Consumer Finances—Purchase of Durable Goods," *Federal Reserve Bulletin*, July, 1958, p. 771.

A frequency distribution should not have too many classes. One of the reasons for grouping data is to reveal the pattern of the distribution. The use of too many classes tends to obscure this pattern.

For example, if we were to use \$100 instead of \$500 class intervals in Example 10.16, the class \$1,500-\$1,999 would be broken down as follows:

\$1,500-\$1,599
\$1,600-\$1,699
\$1,700-\$1,799
\$1,800-\$1,899
\$1,900-\$1,999

We would have only 5 per cent to distribute among these five groups in 1952 and in 1957. This might mean that one or more of these groups would contain 0 per cent. In general, a few items would be spread over many groups, and no pattern would be revealed.

In many cases we must make use of unequal class intervals in order to summarize data adequately. This is especially true for very skewed distributions in which the relative frequencies in the tail are small. Thus, in distributions of family income, we can use an interval as small as \$1,000 for lower incomes because there are many families in such groups. As incomes increase it is necessary to widen the interval to \$2,500, \$5,000, \$10,000, and so on.

At times open-end classes must be employed in order to include a few extremely large or small values without adding excessively to

the number of classes. Such classes either have no upper limit or no lower limit. Thus, in Example 10.16, the classes "under 1,500" and "3,500 and over" are illustrations of open-end classes.

In case of doubt, it is better to have too many classes in a first draft of a table than too few. Smaller classes can be combined into larger classes. However, larger classes cannot be broken down unless the data are retabulated. Thus, in Example 10.15, if we knew the proportion of men who wear small-size shirts, we could not determine the proportion who wear size 14 or size 14½. However, if we knew the proportions for each of the individual sizes, we could combine them into small, medium, large, and extra-large.

A frequency distribution should not have too few classes. Too few classes oversummarize the data and blur the pattern of distribution. The table based on the Survey of Consumer Finances oversummarizes the data for 1957. The "\$3,500 and over" class could have been broken down into additional intervals to give us a better picture of the distribution. However, this would have resulted in undersummarization for the year 1952. To secure uniformity, and thus to facilitate comparisons between the two years, the same classes are used. This emphasizes the fact that the main consideration is always the purpose of the table.

Class intervals should be chosen so that the mid-value of each class interval is approximately equal to the mean of the items in the interval. The mid-value of the class ordinarily may be taken as the average of the lower limits of two successive classes. Thus, the mid-point of the class \$1,500–\$1,999 is equal to \$1,750—i.e., $\frac{1}{2}(1,500 + 2,000)$. The concept of the mid-value as a class average is essential if computations of statistics are to be based upon frequency distributions.

When included as part of a final report, the frequency distribution should be presented as a formal statistical table. As such it should be self-explanatory, clear, and attractive. Any format that accomplishes these ends is satisfactory. If you consult a book that contains statistical tables such as the *Statistical Abstract of the United States*, you will find many acceptable formats. Let's reconsider the table shown in Example 10.16 as a basis for discussing certain essential parts of a statistical table.

First, note that the table has a *title*. The title as a rule answers the questions: "What?," "Where?," and "When?" Each column has a *column heading* which consists of the characteristic (e.g., price) and the unit of measurement (e.g., dollars). The first column

is called the *stub* and lists the classes. *Footnotes* are directly below the table, and the *source* follows the footnotes. Footnotes are needed to explain characteristics of a table that are not self-evident but that are essential to an understanding of the data. The source is the only way the reader can trace the figures to their origin for checking or for obtaining supplementary information. The absence of a source should arouse suspicion about the data in a table.

The table has horizontally ruled lines following the title, after the column headings, between the total figures and the body of the table, and at the close of the table. Vertically ruled lines are used to separate the columns. This method of ruling lines is not mandatory. It has, however, been found that such lines enhance the appearance of the table. They also make for greater clarity and ease of reading.

10.6 Computation of Statistics: The Mean

After the data are edited, coded, and tabulated, pertinent statistics can be calculated. When the sample size is large, it is possible to save time and expense by making these calculations from a frequency distribution. Such calculations are useful especially in the determination of the sample mean and the sample standard deviation. The mean and the standard deviation, as you know, are commonly used criteria for decision. In addition, the sample standard deviation may be used to estimate the population standard deviation which, in turn, may be used to estimate the standard error of the mean.

The methods of computation that we shall consider are most useful when calculations are performed by hand or with a desk calculator, or when secondary data are available only in the form of a frequency distribution. When tabulating machines or electronic computers are available, it is simpler to perform calculations from the raw data without grouping them in a frequency distribution.

In this section we shall consider the computation of the arithmetic mean from a frequency distribution. We assume in the computations that the mean of the items in a class is equal to the value of the mid-point of that class. Thus, in Example 10.17, we assume that the mean of the 6 items in the class \$1.50 to \$1.59 is equal to \$1.55, the mid-value of that class. A similar assumption is made for the other classes in the distribution. If this assumption were true, then the mean of the data grouped into a frequency distribution would be exactly equal to the mean of the ungrouped data. However, this assumption rarely is fulfilled in practice. Nevertheless, if the class

limits are selected carefully, and if the class intervals are not very wide, the error resulting from this assumption will be relatively small.

Example 10.17 A trade association conducts a wage survey. It gathers information on the guaranteed minimum wage of inspectors. The work sheet for the computation of the average minimum wage is shown below:

WORK SHEET FOR COMPUTATION OF MEAN

(1) Wage (in Dollars)	(2) Mid-value of Class m	(3) Number of Workers f	(4) Total Value of Class fm
1.50-1.59.....	1.55	6	9.30
1.60-1.69.....	1.65	11	18.15
1.70-1.79.....	1.75	26	45.50
1.80-1.89.....	1.85	40	74.00
1.90-1.99.....	1.95	32	62.40
2.00-2.09.....	2.05	21	43.05
2.10-2.19.....	2.15	12	25.80
2.20-2.29.....	2.25	8	18.00
2.30-2.39.....	2.35	4	9.40
Total.....		160	305.60

$$\bar{x} = \frac{\Sigma fm}{n} = \frac{305.60}{160} = \$1.91$$

The steps in the computation of the arithmetic mean are indicated in Example 10.17. The first column of the work sheet shows the classes; the second column, the mid-value of each class; and the third column, the frequency in each class. In order to compute the mean it is necessary to determine the sum of all the items. To obtain the sum of the items in each class, we multiply the frequency by the mid-value of the class. Thus, to obtain the sum of the items in the first class, 1.50 to 1.59, we multiply the frequency, 6, by its mid-value, 1.55. The sum of the items for each class is entered into the fourth column—the frequency times the mid-value, fm . The total of this column is the sum of all the items. If we divide this sum by the number of items, we obtain the arithmetic mean. In symbols, we write

$$\bar{x} = \frac{\Sigma fm}{n}$$

The computation of the arithmetic mean may be simplified further by coding the data. To code data we add, subtract, multiply, or divide each data item by a constant. In this manner, it is possible to reduce the magnitude of the data or otherwise simplify them and

thereby facilitate computations. In Example 10.18, we have illustrated the method of computing the mean from coded data. The first step in coding to compute the mean is to subtract a constant from each of the mid-values. If each mid-value is reduced by a constant amount, the mean will be reduced by the same amount. This simplifies the computation but gives us a coded mean. We can decode the data by adding back the same constant. The constant is generally a mid-value near the center of the distribution. Thus, in Example 10.18, 1.95 has been subtracted from each mid-value. The results are shown in column (4).

You will observe that each of the coded mid-values is an exact multiple of the class interval. This always will be true when all class intervals are equal. It is, therefore, possible to introduce an additional simplification by dividing the coded mid-values by the class interval. Thus, in Example 10.18 the coded mid-values are divided by .10. The two steps of the coding operation can be summarized symbolically as follows:

$$d = \frac{m - 1.95}{.10},$$

where d represents the final coded value. The values of d appear in column (5). Thus, for the first class

$$d = \frac{1.55 - 1.95}{.10} = -4.$$

We now find the mean of the coded data. To do so it is first necessary to multiply each coded mid-value (d) by its frequency (f) and divide the sum of the fd values by n . This gives us the coded mean, namely, $-\frac{64}{160}$. To decode the mean we must reverse our coding operations. We, therefore, multiply the coded mean by the class interval (.10) to counteract the earlier division. This gives $-\frac{6.4}{160}$. We then add 1.95 to this result to counteract the effect of the earlier subtraction. Thus, the value of the mean is

$$\begin{aligned}\bar{x} &= 1.95 - \frac{6.4}{160} \\ &= 1.95 - .04 \\ &= \$1.91.\end{aligned}$$

This is exactly the same as the result obtained by direct computation from the uncoded values in Example 10.17.

Example 10.18 The mean of the wage data in Example 10.17 is calculated from coded observations in the work sheet that follows:

WORK SHEET FOR CALCULATION OF MEAN FROM CODED DATA

(1) Wage (in Dollars)	(2) Mid-value of Class m	(3) Number of Workers f	(4) $m - 1.95$	(5) Coded Mid- values d	(6) fd
1.50-1.59	1.55	6	-.40	-4	-24
1.60-1.69	1.65	11	-.30	-3	-33
1.70-1.79	1.75	26	-.20	-2	-52
1.80-1.89	1.85	40	-.10	-1	-40
1.90-1.99	1.95	32	0	0	0
2.00-2.09	2.05	21	+.10	+1	21
2.10-2.19	2.15	12	+.20	+2	24
2.20-2.29	2.25	8	+.30	+3	24
2.30-2.39	2.35	4	+.40	+4	16
Total		160			-64

$$\begin{aligned}
 \bar{x} &= 1.95 + \frac{\Sigma fd}{n} \times (\text{CLASS INTERVAL}) \\
 &= 1.95 + \frac{(-64)}{160} \times (.10) \\
 &= 1.95 - .04 \\
 &= \$1.91
 \end{aligned}$$

To calculate the mean from a frequency distribution, the mid-value of each class must be known. The mid-value of an open-ended class, however, cannot be determined. Thus, it is not possible to compute the average price paid for automobiles by consumers from the data in Example 10.16. Hence, it is important that open-ended intervals be avoided if possible when the mean is needed as a basis for decision. The average value of the items in an open-ended interval should be given in a footnote if such an interval must be used.

10.7 Computation of Statistics: The Standard Deviation

In this section we turn to the calculation of the sample standard deviation. Heretofore, we have computed the sample standard deviation from the formula

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}} \quad (i = 1, 2, \dots, n).$$

This formula has the advantage of pointing up the meaning of the standard deviation, but it is not the easiest to use for purposes of computation. To simplify the computation of s , we shall make use of the fact that

$$\Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} \quad (i = 1, 2, \dots, n).$$

The formula for the standard deviation, thus may be written as follows:

$$s = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n}} \quad (i = 1, 2, \dots, n)$$

$$= \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2} \quad (i = 1, 2, \dots, n)$$

Note that with this method, it is not necessary to compute the deviations about the mean. We need only the values of $\sum x_i$ and $\sum x_i^2$.

Example 10.19 Suppose that we compute the standard deviation of the following 5 observations: 5, 8, 10, 4, 3. The mean of these observations is 6. We now compute s by the use of both methods. The following work sheet provides the basic information:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
5	-1	1	25
8	2	4	64
10	4	16	100
4	-2	4	16
3	-3	9	9
Total 30	0	34	214

The sample standard deviation, from its definition, is

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{34}{5}}$$

$$= \sqrt{6.8}$$

$$= 2.61.$$

We now use the alternative method of computing s .

$$s = \sqrt{\frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2}$$

$$= \sqrt{\frac{214}{5} - \left(\frac{30}{5}\right)^2}$$

$$= \sqrt{42.8 - 36}$$

$$= \sqrt{6.8}$$

$$= 2.61$$

Observe that both methods yield the same result.

A convenient formula for calculating the standard deviation from observations grouped in a frequency distribution is

$$s = \sqrt{\frac{\sum fm^2}{n} - \left(\frac{\sum fm}{n}\right)^2},$$

where m again represents the mid-value of a class and f the frequency in a class.

We can code observations to simplify the computation of the standard deviation in the same manner as for the mean. In Example 10.20, we calculate the standard deviation for the same set of data for which the mean was computed in Example 10.17. Because we ordinarily compute the standard deviation at the same time as the mean, the work sheet illustrates the columns needed to compute both statistics.

You will recall that the first step in coding is to subtract a constant from each mid-value. This operation does not affect the value of the standard deviation because the standard deviation depends only on deviations from the mean. If we reduce each value by a constant, the mean is reduced by the same amount. Thus, the deviations from the mean remain unchanged, and, hence, so does the standard deviation. Because of this, it is unnecessary to compensate for the subtraction of a constant from each observation in computing the standard deviation.

However, when we divide the observations by a constant, the standard deviation is also divided by the same constant. Therefore, it is necessary to multiply the standard deviation calculated from the coded data by that constant in order to obtain the final result.

In the calculations shown in Example 10.20, 1.85 was subtracted from each mid-value (instead of 1.95 as heretofore) in order to show that any mid-value may be subtracted. The results were then divided by the class interval, .10. The coded mid-values are shown in column (5).

Example 10.20 The work sheet for the calculation of the mean and standard deviation for the wage data in Example 10.17 follows:

WORK SHEET FOR THE CALCULATION OF MEAN AND
STANDARD DEVIATION FROM CODED DATA

(1) Wage (in Dollars)	(2) Mid-value of Class m	(3) Number of Workers f	(4) $m - 1.85$	(5) Coded Mid-values d	(6) fd	(7) fd^2
1.50-1.59	1.55	6	-.30	-3	-18	54
1.60-1.69	1.65	11	-.20	-2	-22	44
1.70-1.79	1.75	26	-.10	-1	-26	26
1.80-1.89	1.85	40	0	0	0	0
1.90-1.99	1.95	32	+.10	+1	32	32
2.00-2.09	2.05	21	+.20	+2	42	84
2.10-2.19	2.15	12	+.30	+3	36	108
2.20-2.29	2.25	8	+.40	+4	32	128
2.30-2.39	2.35	4	+.50	+5	20	100
Total		160			96	576

The value of the mean is

$$\begin{aligned}\bar{x} &= 1.85 + \frac{\Sigma fd}{n} \times (\text{CLASS INTERVAL}) \\ &= 1.85 + \frac{96}{160} \times (.10) \\ &= 1.85 + .06 \\ &= \$ 1.91 .\end{aligned}$$

The value of the standard deviation is

$$\begin{aligned}s &= \sqrt{\frac{\Sigma fd^2}{n} - \left(\frac{\Sigma fd}{n}\right)^2} \times (\text{CLASS INTERVAL}) \\ &= .10 \sqrt{\frac{576}{160} - \left(\frac{96}{160}\right)^2} \\ &= .10 \sqrt{3.60 - .36} \\ &= .10 \sqrt{3.24} \\ &= $.18 .\end{aligned}$$

Notice, in particular, that it is necessary only to multiply by the class interval in order to decode the results and obtain s .

10.8 Machine Tabulation and Calculation

In surveys where the number of respondents is large, or where there are many observations to be tabulated, machine tabulation usually is more effective than hand tabulation. It is generally stated that there should be at least 5,000 returns to process before machine tabulation can pay its way. However, this is just a rule of thumb. Thus, if an organization has tabulating equipment and free time available, it might be more economical to use the tabulating equipment even with fewer returns.

Methods of machine tabulation were first used to tabulate the census of population data more than 50 years ago. Since then, tremendous improvements have been made in the machines, but the basic ideas remain the same. Today, there are two kinds of equipment in common use: the Hollerith machines (International Business Machine Corporation equipment) and the Powers machines (Remington Rand Corporation equipment). We shall devote our attention to the International Business Machine (IBM) tabulating equipment.

The first step in machine tabulation is to punch information on cards. Bits of information are represented by various combinations of holes in a card. Figure 10.1 presents an illustration of a punch card. The card has 80 columns and 12 positions in each column where

Data are punched into cards by means of a *key-punch* machine. The cards feed automatically into the punching unit. The holes are punched by depressing typewriting keys for alphabetical information and a 10-digit bank of keys for numerical information. After the cards are punched, the machine stacks them automatically and they are ready for subsequent operations.

The speed with which data can be punched into the cards has been increased through a "mark sensing" process. The holes to be punched are marked with a special pencil. The cards are then put through a machine which punches holes wherever a mark appears.

If cards are punched by hand, the cards may be checked for accuracy by a *verifier*. The operator repunches the data on the verifier, and the machine stops whenever there is a disagreement between the data in the card and the information being repunched. Another method of verification is to have the operator punch a new card and compare the two cards by machine. (See *collator*, below.)

The punched data may be printed on the top of the card by means of an *interpreter*. This machine, guided by the holes, prints the corresponding numbers and letters. However, the newer key-punch machines print the numbers and letters as they are being punched.

The *sorter* is a machine which separates groups of cards, or which arranges cards in a desired order. The sorter drops the card into a pocket that corresponds to the hole punched in a given column. Thus, in Example 10.21, if we wished to separate the "owner-occupied" cards from the "rented" cards, we would sort for column 5. All the "rented" cards would fall in the "1" pocket, and the "owner-occupied" cards in the "2" pocket. A *sorter counter* may be attached to the sorter to count the number of cards in each pocket. Sorters work at a rate of 450 to 1,000 cards a minute. Thus, in about 5 to 12 minutes we could determine the number of owner-occupied and rented dwelling units for 5,000 respondents.

The sorter can also arrange observations in an array. For example, it could arrange rents in order from the lowest to the highest value. This would facilitate the construction of a frequency distribution of rents.

The *tabulator* is a machine that can add, subtract, and print information. Tabulators can print about 100 to 150 lines a minute. They can add about 150 cards a minute if only totals are to be printed. Thus, for the rents in columns 7-9, the tabulator could add and record the total rent paid for apartments with 1 room, 2 rooms, 3 rooms, and so on. It also could record the number of cards in each apart-

ment-size group. It then could print the grand totals for rents and the number of cards. All these operations could be performed in approximately 40 minutes for 5,000 cards.

The *reproducer and summary punch* is a machine that may be hooked up to the tabulator. It punches summary information into a card. Thus, if there were 500 3-room dwelling units with a total rent of \$28,000, the machine would punch this information into predesignated fields of a summary card. The reproducer automatically verifies that the information being punched is identical with the totals in the tabulating machine.

The *multiplier or calculator* is a machine that can multiply or divide numbers punched into two fields. It also can add the numbers in a third field to a product or quotient. Furthermore, it can round off decimals to any place desired. The results are punched into a field that has been set aside for that purpose. In the rental survey illustration, the cards punched on the *reproducer-summary punch* machine could be put through the multiplier to get an average rental for the dwelling units in each class interval, and for all the 5,000 dwelling units. Thus, the multiplier could divide \$28,000 by 500 to obtain \$56 as the average rental for the 3-room dwelling units.

Another useful machine is the *collator*. Let us assume that rents are arranged in array form on two sorters. The collator could merge the cards in both stacks so that the array form is maintained. Also, if we wished to select, from the cards, all rented dwelling units with 3 rooms and rentals over 100 dollars, the collator could perform this operation. It would then replace the cards in their original order.

Electronic computers which combine all the operations of the afore-mentioned tabulating machines have been developed in recent years. These computers also perform other useful operations, all at a tremendous rate of speed. Some computers, for example, can perform as many as 35,000 additions in one second. The speed at which these computers operate has introduced a new era in expediting the flow of data, and in the solution of problems in all fields of endeavor. However, we still are unaware of all the potentialities of computers. Data must be "programmed" before they are processed by a computer. Programming is a plan for processing data. A program consists of a set of instructions that specify the operations a computer has to perform and the sequence of these operations.

The electronic computer requires a longer time for the preparatory cycle in data processing than other methods. It is of utmost importance, therefore, that the planning be done very carefully. To the

extent that planning is poor, or unexpected difficulties occur after processing begins, the computer very easily may magnify rather than reduce the problems of data flow.

As in the case of machine tabulation, data must be prepared in a form that the computer can accommodate. There are three methods of data preparation: (1) punch card, (2) paper tape, and (3) magnetic tape. To record information, holes are punched into cards, holes and/or spots are put on a paper tape, or magnetic spots are placed on a magnetic tape. These holes or spots represent numbers, letters, special symbols, or instructions. The magnetic tape economizes on space since as many as 534 impulses can be marked on an inch of tape. The cards are least economical as space savers.

Cards or tapes are prepared with *input preparation equipment*. The data then are carried into an *input reading mechanism*. The information is translated into electronic equivalents that are fed into the *processing unit*. The processing unit consists of a *memory unit*, a *control unit*, and an *arithmetic and logic unit*.

Data and instructions enter the memory unit first. This is a device for storing the information coming from the input unit, and holds data until they are needed for processing. After data are processed, they may be returned to the memory unit for future use. (In some computers it takes only 12 millionths of a second to find a given number, when needed, in the memory unit.)

The control unit interprets the instructions conveyed by the impulses and directs the sequence of the operations to be followed.

The arithmetic unit can perform the following operations: addition, subtraction, multiplication, and division. The basic operation is addition. Subtraction is performed by the addition of complements. Multiplication is nothing more than successive additions, and division consists of successive subtractions.

The machine also is able to translate final results into decisions. The decision process consists of comparing numbers and indicating some course of action on the basis of the comparisons.

Example 10.22 The U.S. Bureau of the Census is required by law to protect individual respondents to its surveys from the disclosure of their reported figures. In this connection, the Bureau uses the following decision rule: Information (such as manufacturers' sales) is *not* published if the two largest firms account for more than half of the total in a category (e.g., more than half the sales of a product in the United States).

Since the Census Bureau has many categories to which to apply this decision rule on disclosure, it recently has adapted the use of an

electronic computer to the problem. The machine is directed to add values for a category under consideration, to add the values of the two largest firms, and compute the ratio of the latter to the former. If the ratio is .50 or less, the machine prints a symbol indicating that no disclosure is being made. If the ratio is more than .50, the machine indicates the nature of the disclosure, the total for the two largest companies, and their identification numbers.

Heretofore, such checks on disclosure had to be made by clerks, a far more time-consuming system.⁵

The results of the processing operations are reported by means of an *output unit*. The results flow in the form of impulses from the memory unit to the output unit where the results appear as holes in cards, spots and holes on paper tapes, or spots on magnetic tapes. These are used to print the information. Some machines feed the impulses from the output unit directly to a printing device.

The electronic computer uses a language that has only 2 characters in it: "no impulse" which represents 0, and an "impulse" which represents 1. Thus, it is necessary to convert all numbers, letters, and instructions into a binary system: a system which utilizes only a 0 and 1.

Example 10.23 We illustrate the use of the binary system for the digits 0 to 9 and for the letters in the alphabet:

<i>Decimal Digit</i>	<i>Binary Form</i>
0.....	0000
1.....	0001
2.....	0010
3.....	0011
4.....	0100
5.....	0101
6.....	0110
7.....	0111
8.....	1000
9.....	1001

Thus, 4 circuits are needed for each digit. The number 862 would be written as: 1000 0110 0010.

For letters, 6 circuits are used. A is 01 0001, B is 01 0010, C is 01 0011, and so on. Since there are 64 possible combinations of binary digits using 6 circuits, there are more than enough to take care of the 26 letters of the alphabet. The remaining combinations are used to designate certain symbols and punctuation marks.

We have discussed, in an elementary way, some of the basic principles and uses of tabulating machines and electronic computers. The

⁵ Maxwell R. Conklin and Owen C. Gretton, "Some Experience with Electronic Computers in Processing the 1954 Census of Manufactures," *The American Statistician*, Vol. XII, No. 4 (October, 1958), pp. 19-22.

proper use of these machines, as you can see, will facilitate the flow of information in statistical studies. Nevertheless, their use makes the problem of planning for data flow exacting. Plans must be clearly and carefully made if machines are to aid in the decision-making process. Remember that the machines themselves can do no more than they are directed to do.

10.9 The Final Report

The final step in a statistical study is to record the results in the form of a *final report* for future as well as for immediate reference. The most carefully planned and executed study is wasted unless the results are conveyed to the people responsible for carrying out and supervising the action indicated by the study. Furthermore, a record of the experiences of the study should be kept for reference in similar studies which may be conducted subsequently.

Although each problem situation calls for its own special type of report, and although each report presents its own special requirements, there are some general aspects of report writing that we may discuss. First of all, the final report should usually appear in two parts. The first part should summarize, in a nontechnical way, the conclusion—the statistical findings and the action indicated by these findings. The second part should be devoted to a more detailed explanation of the more technical aspects of a study, including a description of the sampling procedure, the risks involved, and so forth.

The nontechnical part of a report may assume many forms. It may vary from an oral communication or brief memorandum to a formal published document. As a rule, reports of studies conducted within an organization to solve purely internal problems are brief and informal. However, reports of studies conducted by outside consultants assume a more formal character. Similarly, in large organizations, methods of reporting may be highly formalized. The results of secondary studies, especially those designed for wide use, generally are presented as formal publications.

As a minimum requirement, even the most informal nontechnical part of a report should contain a statement of the problem and a summary of the findings. Thus, a report on an estimation problem should include the reason for the study and the resulting estimates. A report on a testing problem should contain a statement of the purpose of the study, the sample results, and the appropriate action indicated by these results. Thus, an inspection report on whether to accept or to reject a lot should contain the nature of the material in-

spected, the sample size, how the sample was drawn, the number of defectives in the lot, and the action to be taken on the lot. To facilitate reports on such repetitive studies, special forms may be designed.

The second part of a report should include additional, more technical, information concerning a study. This part of a report also will vary with the nature of the statistical study and the problem itself. Such information is particularly important for secondary data, which were not collected under the auspices of those using the results for a decision. Here we shall list the type of information that should be checked in the report of any secondary study. At the same time, you should recognize that it also is necessary to include each of these requirements in a report of the results of a primary study (which, after all, also may at a later time and for another problem become a secondary study).

Hence, the technical section of a final report should include remarks on the following eleven aspects of a study:

1. *Statement of the purposes of the study.* Secondary studies are usually not directed at answers to a specific problem. Therefore the consumer of such studies should be informed on why the study was undertaken and on the ways in which it is expected that the results of the study may be utilized.
2. *Description of the population.* A precise description should give the geographical region and the categories of material covered by the study. It should describe the frame that was utilized. Care must be taken that there is no misunderstanding about the coverage of the study.
3. *Nature of information collected.* The report should show the information collected in sufficient detail. Information collected but not reported on should also be mentioned. Thus, hourly wages might be summarized as "\$2.50 and over," but if detailed figures of actual wages were collected, this point should be mentioned. If possible, the schedules used to collect the data should be reproduced, or made available upon request.
4. *Method of collecting data.* For example, a report should indicate whether personal interview, telephone interview, mail questionnaire, panel, or a combination of these methods was used.
5. *Method of sampling.* The method of sampling should be described. The sample size, as well as the arrangements for call-backs in case of nonresponse, should be given. Also the way in which the sample size was determined should be discussed.

6. *Decision rule.* The decision rule should be given in full, together with a description of the manner in which it was determined.

7. *Repetition.* Is the study a one-shot operation, or is it one in a series of similar surveys? If the latter, then differences between this and earlier surveys should be pointed out.

8. *Date and duration.* This includes the period of time to which the data pertain.

9. *Assessment.* A statement should be made of the extent to which the study actually fulfilled its intended purposes.

10. *Responsibility.* The name of the organization sponsoring the study, as well as the names of the people conducting it, should be stated. This information often is a rough indicator of the quality of a study.

11. *Accuracy.* A general indication should be given of the degree of accuracy attained. If one is to base decisions on the results of a study, he must know the magnitude of the errors involved. The magnitude of the sampling error, as well as the possible sources of procedural biases or nonsampling errors, should be indicated. If possible, an estimate of the magnitude of the biases should be given.

Every bona fide study should contain information of the type described in this section. One should view with suspicion statistical studies which do not include such information in their reports and whose sponsors refuse to furnish such facts upon request.

10.10 You Should Now Know That

Data flow deals with the flow of information from the time that the sample of elementary units is chosen to the time that the observations are presented in final form to serve as guides for decision.

There are many methods of data collection including direct observation, the personal interview, the mail questionnaire, telephone questioning, and mechanical devices.

The choice of the method of data collection depends upon the nature of the problem, the population of study, costs, and time considerations.

The organization of the collection procedure is important in small-scale and in large-scale studies.

Processing the data requires that they be edited to determine their completeness, consistency, and accuracy. It also requires the coding of observations to facilitate their tabulation.

The tabulation of data involves the summarization of results in statistical tables.

The form of a frequency distribution is determined by the purpose it is to serve.

It sometimes is possible to save time and expense by computing statistics from a frequency distribution.

It may be more efficient to compute statistics from coded observations than from original observations.

In large-scale studies, machine tabulation is more economical than hand tabulation.

Electronic computers, if utilized efficiently, can be of considerable use in facilitating the flow of information in a statistical study.

A final step in a statistical study is the recording of the results in a final report.

10.11 You Should Now Be Able to Solve

1. Critically and constructively discuss the following statement:

All of the steps in data flow including collection of data, editing, coding, tabulating, and computing are essential parts of every statistical study. Therefore, each of these steps requires a 100 per cent verification of the accuracy of the work done.

2. Instructions often are mailed to interviewers participating in nation-wide market research studies. What are the possible shortcomings of this procedure?

3. For each of the following situations, discuss which method of data collection (personal interview, mail questionnaire, etc.) would be preferable.

- a) A study by the federal government to determine the value of retail sales each month.
- b) A survey by a trade association to determine the attitude of its members toward the publication of a trade journal.
- c) A study by the personnel department of a firm to determine attitudes of employees toward a contemplated plant cafeteria.
- d) A study to determine the attitude of physicians in New York City toward a recently marketed antibiotic.
- e) A study by a local real estate firm to determine the number of homes for sale in the area.

4. The decision rule for an inspection procedure is as follows: Select a simple random sample of 100 pieces. If 3 or less are defective, accept the lot; otherwise reject the lot. The number of defectives,

found in the samples selected from the first 50 lots inspected, was tabulated as follows by the inspector:

<i>Number of Defectives in Sample</i>	<i>Number of Lots</i>
0.....	3
1.....	5
2.....	15
3.....	25
4.....	0
5.....	1
6.....	<u>1</u>
Total.....	50

Do these results appear suspicious? Explain.

5. In reply to a questionnaire a company reports the following ratios for the year 1958:

Cost of goods sold/sales.....	.78
Administrative expense/sales.....	.08
Selling expense/sales.....	.15
Net profit before taxes/sales.....	.05

Would you question the accuracy of this return? Explain.

6. The XYZ Corporation has 1,018 accounts receivable. It selects a sample of 100 accounts to estimate the average value of its accounts receivable as at December 31, 1958.

The 100 sample observations (in dollars and cents) are as follows:

15.00	5.60	44.08	81.65	8.85	20.50	25.75	35.64	36.87	65.85
88.37	43.07	63.29	35.70	36.19	17.08	76.40	45.50	49.61	41.85
51.10	25.30	22.25	45.50	37.63	53.36	47.84	42.56	23.50	70.65
46.24	56.34	44.80	50.04	55.00	47.42	55.10	52.09	49.16	39.70
16.10	12.80	40.75	13.90	46.36	27.75	48.29	26.25	52.81	31.80
40.15	74.60	59.05	43.70	34.83	51.30	62.00	50.60	42.56	42.63
48.94	56.43	42.35	64.80	54.07	2.30	33.35	45.90	65.10	38.40
53.71	28.65	67.80	57.23	42.97	66.30	30.90	34.15	58.15	29.50
43.61	41.25	45.15	43.38	32.70	44.46	68.90	19.16	69.70	59.95
54.60	79.35	10.50	60.28	37.39	85.50	33.20	58.85	24.70	46.10

- Summarize the data in a frequency distribution. Use a \$10 class interval, with \$0.00 to \$9.99 as the first class.
- Present the frequency distribution as a formal statistical table for use in a final report.

7. Compute the mean and the standard deviation of the sample from the frequency distribution in Problem 6.

What would be your estimate of the average value of accounts receivable in the population?

Compute the standard error of the estimate.

Estimate the total value of the population of accounts receivable and compute the standard error of this estimate.

10.12 You Will Also Find That

The following references discuss many of the aspects of data flow, including methods of data collection, organization of the collection procedure, and methods of processing and tabulating data, with particular reference to studies of human populations:

DEMING, W. EDWARDS. *Some Theory of Sampling*, pp. 1-23. New York: John Wiley & Sons, Inc., 1950.

PARTEN, MILDRED. *Surveys, Polls, and Samples: Practical Procedures*. New York: Harper & Brothers, 1950.

Almost every statistics textbook discusses the construction of frequency distributions and methods of computing statistics from such distributions. For example, you might consult:

CROXTON, FREDERICK E., and COWDEN, DUDLEY J. *Applied General Statistics*, pp. 153-70, 173-82, 215-22. 2d ed. New York: Prentice-Hall, Inc., 1955.

MILLS, FREDERICK C. *Statistical Methods*, pp. 40-64, 89-94, 116-20. 3d ed. New York: Henry Holt & Co., 1955.

NEISWANGER, WILLIAM A. *Elementary Statistical Methods*, pp. 224-50, 256-66, 311-18. Rev. ed. New York: Macmillan Co., 1956.

A memorandum entitled "Recommendations concerning the Preparation of Reports on Statistical Surveys" was prepared by the United Nations Subcommittee on Statistical Sampling in 1948. The memorandum is obtainable from the United Nations Statistical Office and also is reprinted, in part, in:

YATES, FRANK. *Sampling Methods for Censuses and Surveys*, pp. 141-44. New York: Hafner Publishing Co., 1949.

Continuing Decision Problems and the Control Chart

11.1 Introduction

In this chapter we shall consider the application of statistical theory in continuing decision problems. A continuing decision problem, as the name implies, is a problem situation which requires almost unceasing attention—one which calls for frequent decisions on whether or not to take a certain course of action.

Example 11.1 An investor usually faces a difficult problem when he decides whether or not to purchase a particular security. It is a problem that can be rationally resolved only after carefully investigating a company.

However, when an investor makes a decision to purchase some security, he should recognize that he will face an even more exacting problem—a continuing decision problem on whether to retain or to sell the security. This is, obviously, a persistent and unceasing problem which calls for almost continuous attention, at least until the security is sold. As one large brokerage house has advertised:

“ . . . no investment decision can ever be a final one. Changed conditions constantly create changed investment opportunities and the investor must constantly appraise them . . . every day that you own and hold 50 shares of Typical Manufacturing, you say in effect: ‘I’m satisfied it’s the best investment I can make of my money.’ ”
(Merrill Lynch, Pierce, Fenner & Smith.)

There are many business situations which call for continuing decisions. A commonly occurring problem requires a decision on a process or a system in one’s company—a decision on whether to modify the process or to allow it to continue operating as is.

Example 11.2 Consider any manufacturing company which sets up a process to produce a product. It is important to design such a process so that it is capable of producing a relatively uniform, economical, and useful product. Hence, the design, development, and introduction of the production process involve important decisions.

However, once a process is introduced, an equally important problem arises. It is caused by the recurring possibility that the quality level of the process is deteriorating. Hence, the following question must be answered at regular intervals: "Should we let the process continue to run as is, or should we examine and possibly modify it?"

Example 11.3 A personnel administrator of a large company periodically considers a review of his subordinates' personnel decisions (hiring, firing, promotions, demotions, ratings, etc.). The purpose of this review is to modify, if necessary, any trouble spots in personnel operations. Thus, the administrator faces a continuing decision problem on whether to let his department continue to operate as heretofore, or to review recent decisions as a basis for making some modification in departmental operations.

Other continuing decision problems on processes or systems arise in the inspection of clerical work, in maintaining a safety program, in maintaining inventories, in sales management, in controlling shortages, in records management, and so forth. Many of these problems are statistical, for ordinarily the wise choice of an alternative course of action depends upon knowledge of some pertinent numerical information.

Continuing decision problems which are to be treated statistically obviously require not one sample study but a series of sample studies. On the basis of each sample, a decision on whether or not to take action on a process is made. A particularly simple way of depicting a decision rule and the results of the series of samples for this type of problem is called a *control chart*. In this chapter we shall center our attention on the development of control charts for continuing decision problems.

First, however, a few comments on the historical developments of these statistical techniques should be instructive. As early as 1924, statistical methods were applied to the study of production processes by Dr. Walter A. Shewhart of the Bell Telephone Laboratories. In 1931, Dr. Shewhart published his book¹ on these methods, and this book set the pattern for the subsequent developments in the field. Prior to World War II most applications of quality control to pro-

¹ *The Economic Control of Quality of Manufactured Product* (New York: D. Van Nostrand & Co., Inc., 1931).

duction processes were found in a few plants in the electrical manufacturing and textile industries and in some government arsenals. However, since that time there has been a rapid expansion in the use of statistical quality control in all types of industries. Furthermore, the use of these same statistical methods has been extended from production process control to the control of various administrative and clerical operations.

In contrast, prior to the introduction of statistical techniques to check on production processes, a firm usually became aware of defects in its operations only after customers complained, or only after excessive waste, scrap, and rejects, or a sharp rise in rework time, or the loss of sales to competitors had occurred. In other words, action on the process was taken only after damage to the product and perhaps to the firm had taken place. Furthermore, when product deterioration was discovered, haphazard trial-and-error methods were used to take action on the process. The introduction of Shewhart's statistical methods replaced the haphazard procedure of the past with a scientific, objective, and simple and economical continuing decision procedure. It is now possible to detect deterioration in the process before the end product has become defective.

11.2 Process Variability

Many continuing decision problems, as noted, deal with the question of whether or not to take action on a process. Hence, they require analytical studies, and the population under consideration should be regarded as infinite. In addition, the population of interest is studied not once but at regular intervals.

To illustrate the nature of the problem, consider a firm which produces a film-coated wire. The aim of the firm is to keep the thickness of the coating uniform. Complete uniformity, however, is not possible for variability is a rule of nature. Hence, the population of thicknesses of the film coating must be characterized by some variability. Thus, from hour to hour the thickness of film on the wires will vary simply because of the variability inherent in the process. If, in addition, the nature of the process itself changes in some way (that is, the process shifts), there will be additional variation from hour to hour in the thickness of the coatings. Since some variability is to be expected, a question continuously arises as to whether or not any observed change in the quality characteristic may be disregarded as inherent in the process. The question must be answered because if an observed change is actually due to a significant process change, the process should be investigated.

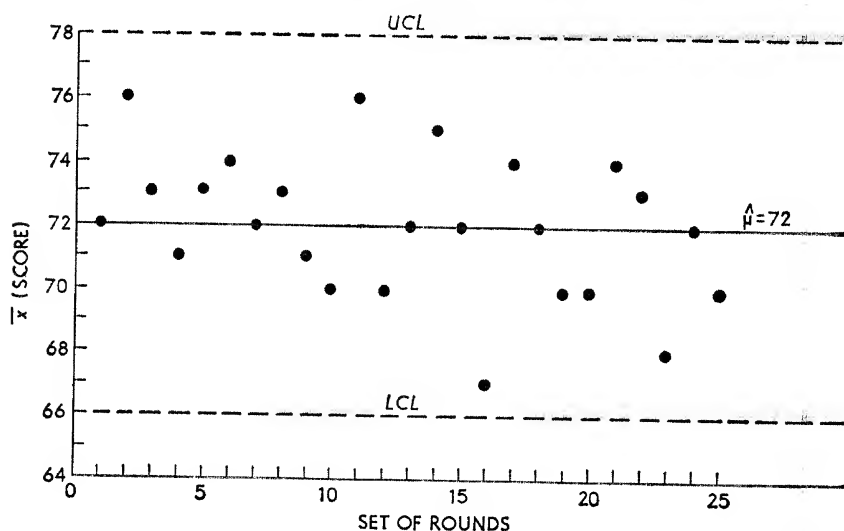
To introduce the statistical methods used to analyze the causes of process variability let us consider a simple illustrative situation.

Example 11.4 A friend of the authors is a professional golfer and keeps a record of his golf scores. Recently, he became displeased with his performances. He had a feeling that something had "gone wrong" with his game and thought his game should be analyzed by another professional golfer. He consulted the authors for an opinion. They grouped the golfer's scores for the last 125 rounds played (on his home course) into 25 groups of 5 consecutive rounds, and computed the mean of each group. The group means are plotted in Figure 11.1.

Figure 11.1 is a *control chart*. The solid horizontal line represents the average score for the 125 rounds played. The two dotted lines that are labeled *UCL* (upper control limit) and *LCL* (lower control limit) are drawn at a distance of 3 standard errors of the mean from the average score of the 125 rounds.

An examination of the chart led the authors to conclude that an analysis of their friend's game might not result in any improvement. How was this conclusion reached? The answer to this question will be postponed. However, after reading the discussion that follows, the answer should be apparent to you.

Figure 11.1
CONTROL CHART FOR MEAN SCORE OF FIVE ROUNDS OF GOLF



You will observe that the means of the sets of 5 rounds vary. We can consider golf playing as a process. The population consists of the infinite number of scores for 18 holes of golf that the golfer is capable of producing. Each group of 5 scores is, therefore, a sample from such a process.

Let us now consider, in a general way, why the golf scores vary. The causes for this variability can be classified into two broad categories: (a) common causes, and (b) special causes. Instead of the terminology "common-special" causes, statisticians also employ the terms "unassignable-assignable," "uncontrollable-controllable," and "chance-significant" causes. We shall use these terms interchangeably.

Special causes of variability are those that bring about changes in the inherent nature of the process. Since this is so, they can usually be identified, and, hence, remedial action may be taken to remove them. Thus, the golfer might not be following through on his swing regularly; he might be shifting his grip; his stance might vary from swing to swing. All of these factors would tend to introduce variability in the golfer's game. These factors, however, could all be detected by a professional golfer, and attempts could be made to correct them.

But, even if every swing were in keeping with sound practice, the scores would still vary from round to round. They would vary because of the common causes influencing the golf-playing process. Common causes may be subdivided into two groups: variable and constant. The variable causes usually consist of a host of factors whose influences are unidentifiable. No doubt there are very many such factors, and the effect of any one of them must be very slight. In fact, the influence is so slight that if one factor were removed, its absence would not be felt. If an additional factor were added, its presence similarly could not be detected. The aggregate effect of these variable common factors, however, is to introduce a pattern of variability into a process.

The constant common causes have a uniform impact on each unit (round of golf). Thus, an injury during childhood might have resulted in the favoring of certain muscles by the golfer when he swings. This might weaken the power of each drive and might result in a constant effect of adding 2 strokes to each round played. Such constant factors, since they affect each round uniformly, obviously do not influence the variability of the scores. They do affect the average or level of the scores. Thus, if the cause of the constant effect of 2 strokes per round were discovered and corrected, the average score would drop from 72 to 70. But, the variability about the average would remain the same. These constant common causes are inherent in a given process and are generally difficult to identify, especially if the process was efficiently designed.

What is the relationship between the cause systems making for

process variability and the decision-making process? Obviously, if the variability of a process is such that its behavior can be attributed to common causes, it is best to leave the process alone. The process is simply operating within the limits of its inherent capability. To search for special causes under such circumstances would be costly and wasteful, and might even result in injecting trouble into a process in which no trouble exists. However, if the behavior of a process is such that the presence of special causes is revealed, then action to look for trouble should be taken.

Since the decision on the process is to be based on sample observations, risks of incorrect action exist. The nature of the possible actions and risks involved in decisions on process variability are as follows:

Action	State of the World	
	No Trouble Exists (Common Causes Affect Process)	Trouble Exists (Special Causes Affect Process)
Leave process alone.....	Correct	Incorrect
Investigate process.....	Incorrect	Correct

The control chart shown in Figure 11.1 presents graphically a decision rule for determining the appropriate action on the process. The chart is interpreted as follows: If sample means fall between the lower (*LCL*) and upper (*UCL*) control limits, leave the process alone; if any sample mean falls on the control lines or outside of the control lines, investigate the process. Since the golfer's average scores fell within the control limits, it was decided that no action need be taken. In other words, it was concluded that the variability in his game was due to common causes—i.e., his game was in *statistical control*. The variability in the scores was attributable to the inherent nature of his playing—and nothing had “gone wrong” with his game.

11.3 The Control Chart for the Proportion Defective

We shall now direct our attention to the construction of the most common types of control charts. In this section we shall consider one of them—the control chart for the proportion defective (called a *p-chart*). The sample proportion of defectives is used as a basis for decision on a process when the quality characteristic under consideration is described as an attribute—usually as defective or nondefective. Such situations occur, for example, when radios are checked

to see whether or not a wire is soldered to the proper binding post; when books are examined to see whether or not any pages are missing; and when attendance records are checked to see whether or not an employee is absent.

To facilitate the succeeding discussion, let us consider the following illustration.

Example 11.5 A manufacturer of automatic dryers decides to use a control chart to check on the process of shaft and washer assembly—one stage in the manufacture of his product. Each sampled assembly is classified defective or nondefective. Hence, a p -chart is called for.

Table 11.1 presents the proportion or fraction defective shaft and washer assemblies during the month of April in samples of 1,000. (It

Table 11.1

**INSPECTION RESULTS FOR SHAFT AND
WASHER ASSEMBLIES**

April, 1958

(1,000 Assemblies Inspected Each Day)

Date	Number Defective	Proportion Defective
April 1.....	28	.028
2.....	42	.042
3.....	36	.036
4.....	40	.040
5.....	35	.035
6.....	37	.037
8.....	33	.033
9.....	41	.041
10.....	40	.040
11.....	44	.044
12.....	39	.039
13.....	42	.042
15.....	45	.045
16.....	28	.028
17.....	29	.029
18.....	25	.025
19.....	31	.031
20.....	43	.043
22.....	60	.060
23.....	37	.037
24.....	76	.076
25.....	31	.031
26.....	30	.030
29.....	61	.061
30.....	47	.047
Total.....	1,000	

is customary to use the term "fraction defective" rather than "proportion defective" in discussions of process control.)

Note that even if the 1,000 assemblies examined on any day were to constitute the entire day's output, they would still constitute a sample. Process control involves analytical studies, and a day's output is only a sample from the infinite population of assemblies under consideration.

The central line of a p -chart is set at π , the process proportion defective. The upper and lower control limits are placed at a distance of 3 standard errors of a proportion above and below the central line, respectively. These limits are, therefore, referred to as *3-sigma limits*. If a sample proportion falls within the 3-sigma limits, it is indicative of the result of common causes and no action is taken on the process. If a sample proportion falls on or beyond the 3-sigma limits, action is taken to look for special causes. Since this decision is based on sample observations, risks of incorrect action exist—risks that we may measure, and that we shall discuss below.

To determine the control limits it is necessary to know the population proportion. Although the process proportion defective is not known exactly, it can be estimated on the basis of the 25,000 observations gathered in the 25 samples. In setting up control charts it has been found, from experience, that 25 samples, as a rule, are needed to obtain a sufficiently precise estimate of the decision-parameter. However, if it is too costly to wait until sufficient data have been gathered, fewer samples may be used to construct the initial control chart.

We know that an unbiased estimator of the population proportion is

$$\hat{\pi} = p = \frac{\text{NUMBER OF DEFECTIVES}}{\text{TOTAL NUMBER OF ITEMS TESTED}}.$$

If we apply this to the data in Table 11.1 we find

$$\begin{aligned}\hat{\pi} &= \frac{1,000}{25,000} \\ &= .040.\end{aligned}$$

We can now compute an estimate of the standard error of the proportion for samples of size 1,000 from the estimate of π :

$$\begin{aligned}\sigma_p &= \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \\ &= \sqrt{\frac{(.040)(.960)}{1,000}} \\ &= .0062.\end{aligned}$$

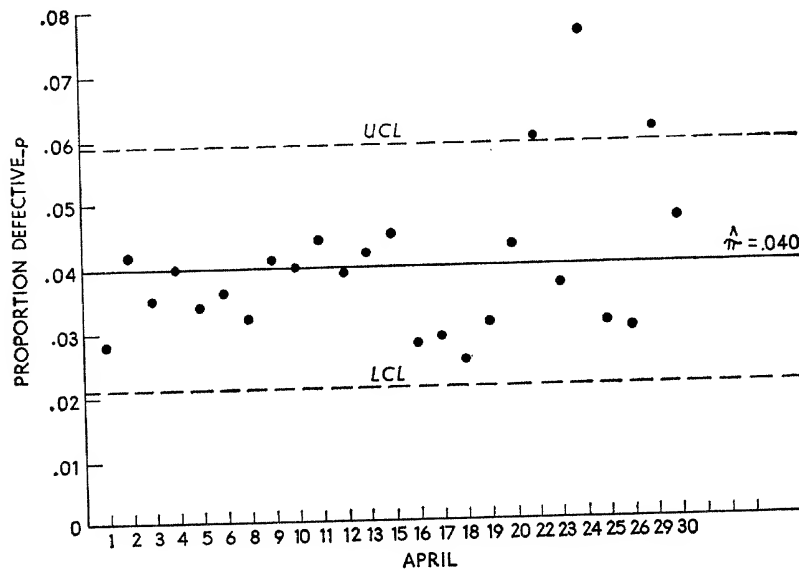
The upper and lower control limits can now be calculated.

$$\begin{aligned} LCL &= \hat{\pi} - 3\hat{\sigma}_p & UCL &= \hat{\pi} + 3\hat{\sigma}_p \\ &= .040 - 3(.0062) & &= .040 + 3(.0062) \\ &= .021 & &= .059 \end{aligned}$$

Figure 11.2 represents the control chart—the graphic representation of the decision rule. The possible values of the sample proportion defective are indicated on the vertical axis, and the samples are identified on the horizontal axis by the date on which they were selected. Horizontal lines are drawn to represent the process average and the upper and lower control limits. The observed sample proportions are plotted above the appropriate dates.

Figure 11.2

CONTROL CHART FOR PROPORTION DEFECTIVE FOR
SHAFT AND WASHER ASSEMBLIES



We now must check to see if all our sample proportions fall between the upper and lower control limits. If they do, then our conclusion is that the process variability is due to unassignable causes and we do not look for any trouble. In our illustration, however, the samples taken on April 22, 24, and 29 indicate the presence of assignable causes. The control chart, therefore, indicates that the process is not in statistical control. Before looking for trouble, however, it is desirable to check the inspection reports to make sure that no

clerical errors were made in reporting the number of defective assemblies.

We have, thus, observed one use of the control chart, i.e., to determine whether or not a process is in statistical control. The process in our illustration is not in control. If a process is not in control, it is not possible to predict the pattern of its future behavior, except to say that it is likely to stay out of control. The control chart will be descriptive of the behavior of sample proportions only if the process is acted upon by common causes. Therefore, in order to use the control chart for maintaining control it is first essential to secure statistical control for the process.

A check must be made to determine the nature of the special causes that seem to be affecting the process and to correct them, if possible. It is also necessary to drop the observations for the days when the process was out of control and re-estimate the process proportion defective. We are interested in the proportion for a process in control. Data based on observations when the process was not in control would distort the results.

Example 11.6 The observations for April 22, 24, and 29 have been excluded, and the process proportion defective has been re-estimated below:

$$\begin{aligned}\hat{\pi} &= \frac{803}{22,000} \\ &= .037.\end{aligned}$$

The revised standard error of the proportion now is

$$\begin{aligned}\hat{\sigma}_p &= \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} \\ &= \sqrt{\frac{(.037)(.963)}{1,000}} \\ &= .0060.\end{aligned}$$

The revised control limits are

$$\begin{aligned}LCL &= \hat{\pi} - 3\hat{\sigma}_p & UCL &= \hat{\pi} + 3\hat{\sigma}_p \\ &= .037 - 3(.0060) & &= .037 + 3(.0060) \\ &= .019 & &= .055.\end{aligned}$$

In Figure 11.3 the revised control chart is drawn and the sample proportions are plotted. You will observe that now all sample proportions fall between the control limits, and we, therefore, conclude that the process was, for those days, in control.

Note again that it is essential for a process to be in control before we can use the control chart as a guide for future action on the proc-

ess. If the process is out of control, the causes for the excessive variability must be sought out, must be corrected, and the process must be brought into control. Statistical theory enables us to predict the future behavior of a process only if it is a random process. It tells us that the same percentage of results will occur within given limits as long as the chance system prevails.

How do we use a control chart, once the limits have been set?

Example 11.7 Let us assume that during May, 1958, 1,000 assemblies were inspected daily with the following results:

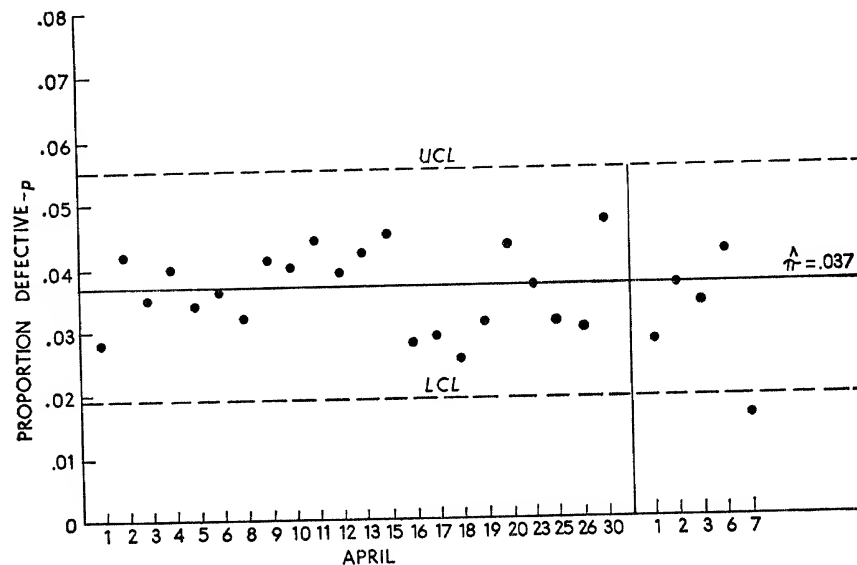
Date	Fraction Defective
May 1.....	.028
2.....	.037
3.....	.034
6.....	.042
7.....	.016

You will observe in Figure 11.3 that for the first 4 days the observed proportion defective fell within control limits. This indicates that the process seems to be in statistical control and the action is to leave the process alone. On May 7th, however, the proportion defective is below the *LCL*. This indicates the possible presence of an assignable cause. The action is to look for the cause of the trouble.

In general, therefore, our rule for the continuing decision problem

Figure 11.3

CONTROL CHART SHOWING REVISED LIMITS FOR PROPORTION DEFECTIVE SHAFT AND WASHER ASSEMBLIES



is: *If the observed proportion defective falls between the control limits, leave the process alone. If the observed proportion defective falls on or outside either limit, look for assignable causes.*

The question as to why we set lower control limits for the proportion defective may have occurred to you. Why do we look for trouble, as in the above illustration, when our process behaves as if it were significantly better? In the first place, when a point is outside the control limits, it indicates that assignable causes are operating on the process. The pattern of variability is no longer random. Upon investigation it might be discovered that the unassignable cause that made for a significant decrease in the proportion defective might at a later time result in a sharp increase in the proportion defective. In the second place, if the assignable cause is found to be a factor that will make for improved production, then the process may be modified to incorporate that factor. In either event the process must be brought under statistical control before we can continue to use the control chart as a basis for action.

The preceding discussion has stressed industrial applications of the p -chart. There are, however, many administrative applications of the p -chart. Two illustrations follow:

Example 11.8 Alden's Inc., is a mail-order house that has been applying control chart techniques to its operations for many years. One such operation involves the process of filling orders in one of the merchandise departments. Every item that is ordered by a customer is selected from stock on the basis of its catalogue number, color, size, and quantity. After the items are selected, they are checked against the order and placed on a conveyer belt to a gravity chute. Several times during the day 100 work units are selected at random from the chute and inspected. A work unit is considered to be incorrectly processed if any of a set of specified error possibilities including catalogue number, color, size, quantity, and price are in error. The proportion of incorrect units is posted on a control chart as quickly as possible. When a point is out of control, remedial action is taken immediately.²

Example 11.9 The control chart technique is used by United Air Lines to control the accuracy of recording plane reservations.

Space on United Air Lines planes is controlled at the Denver office. About 10,000 telephone messages are received daily. The incoming telephone messages are recorded manually on message slips. The telephone wires are tapped 3 times every day, and 200 consecutive messages are recorded. A carbon copy of the message slips is then checked against the recording. The slip may be in error with respect to flight

² James M. Ballowe, *Statistical Quality Control of Clerical and Manual Operations*, Reprint No. 10, Fifth Midwest Quality Control Conference, 1950.

number, date, stations involved, number of seats, or the recording of information that did not enter into the telephone conversation. If one or more of the possible errors occur, the slip is counted as defective. The control charts employ 3-sigma limits, and the process average is 0.5 incorrect messages per 100 transcribed.³

11.4 Control Charts for the Mean and the Range

When a quality characteristic such as the diameter of a cylinder, the tensile strength of a piece of steel, or the resistance of a coil can be measured, the mean may be used as a basis for action on a process. The mean alone, however, is not sufficient to indicate the quality of such processes.

Example 11.10 Two samples each consisting of 4 cylinders are drawn, and the inner diameter (in inches) of each of the 8 cylinders is measured with the following results:

<i>Sample 1</i>	<i>Sample 2</i>
1.52	0.50
1.49	1.10
1.48	2.30
<u>1.51</u>	<u>2.10</u>
$\bar{x}_1 = 1.50$	$\bar{x}_2 = 1.50$

Thus, both samples yield means of 1.50 inches, and both might indicate that the process level is in control. However, the variability for sample 2 is much greater than for sample 1. The cylinders in sample 2 are all either too small or too large, even though their average inner diameter indicates control.

It is, therefore, necessary to use a control chart for the mean ($\bar{x}$ -chart) in conjunction with a control chart for process variability. The criteria for variability that are used are the range (R -chart) or the standard deviation (s -chart).

The two charts, one for the level and the other for variability, used in conjunction, supplement each other in searching for assignable causes. Changes in process level usually require different corrective actions from changes in process variability. Shifts in process means reflect changes in all pieces made. Such factors as a gradual increase in temperature, tool wear, different methods used by workmen on different shifts, or a new batch of material of greater hardness might result in significant shifts in the process mean. Significant changes in process variability, on the other hand, might be traced to such factors as worn bearings, variability in pieces received

³ J. S. Brinkman, "United Air Lines Speeds Reservations," *Paper-Work Simplifications*, No. 16, 1949, pp. 9-10.

from a previous operation, a loose part, or a lack of concentration on the part of an operator. In general, the factors that affect all items result in shifts in the process level, while the factors that affect process variability are operative intermittently and, hence, affect some items and not others.

Let us now consider the construction of an $\bar{x}$ -chart for the process described in the illustration that follows.

Example 11.11 A cannery fills No. 303 cans with corn by a machine process. It wants to set up a control chart so that the drained weight in the cans is at a given average level. Before the control chart can be used as a guide for future production control, we must determine whether the process is operating under statistical

Table 11.2
DRAINED WEIGHT OF CONTENTS OF SIZE NO. 303 CANS
OF GOLDEN SWEET WHOLE KERNEL CORN
(Weight in Ounces)

[illegible]

control. That is, do the drained weights of corn vary because of unassignable causes? Therefore, at 5 time periods during the day a sample of 5 cans that have been filled and sealed are taken from the production line. The cans are emptied, drained, and the weight of the solid contents recorded to the nearest tenth of an ounce. This process is repeated for 5 days to obtain the needed 25 sample observations for setting up the initial control chart. The observations are recorded in the Table 11.2.

The central line in the $\bar{x}$ -chart is set at an estimate of μ —the process average. The upper and lower control limits are the customary 3-sigma limits—in this instance $3\sigma_{\bar{x}}$ since the sample mean is the pertinent statistic.

To determine the control limits, it is necessary to estimate the process average and the process standard deviation from the observations obtained in the 25 samples. The estimate of the process average is the average of all the sample observations. Since all samples are of the same size, we may simply average the 25 sample means.

$$\hat{\mu} = \frac{\Sigma \bar{x}}{m},$$

where m is the number of samples. Hence,

$$\begin{aligned}\hat{\mu} &= \frac{425}{25} \\ &= 17.0 \text{ ounces.}\end{aligned}$$

The standard deviation of the population is usually estimated from the mean of the 25 sample ranges. In Example 7.9 we saw that an estimate of the population standard deviation is given by

$$\hat{\sigma} = \frac{\bar{R}}{d_2},$$

where R is the average of sample ranges and d_2 a constant which depends on sample size. (Values of d_2 for selected sample sizes appear in Table 11.3.)

Table 11.3

SELECTED FACTORS USED IN CONTROL CHART
CONSTRUCTION FOR SAMPLE SIZES 2-7

Sample Size	d_2	A_2	D_3	D_4
2.....	1.128	1.880	0	3.267
3.....	1.693	1.023	0	2.575
4.....	2.059	0.729	0	2.282
5.....	2.326	0.577	0	2.115
6.....	2.534	0.483	0	2.004
7.....	2.704	0.419	0.076	1.924

The estimate of the standard deviation for the canning process is

$$\begin{aligned}\hat{\sigma} &= \frac{2.82}{2.326} \\ &= 1.21 \text{ ounces.}\end{aligned}$$

The sample range rather than the sample standard deviation is often used to estimate the population standard deviation because the range is much simpler to compute. Furthermore, when the sample size is small, the loss in precision of the estimate of the population standard deviation from sample ranges is relatively slight.

The upper and lower control limits for the $\bar{x}$ -chart are:

$$LCL = \hat{\mu} - 3\hat{\sigma}_{\bar{x}} \qquad UCL = \hat{\mu} + 3\hat{\sigma}_{\bar{x}}.$$

But,

$$\begin{aligned}\hat{\sigma}_{\bar{x}} &= \frac{\hat{\sigma}}{\sqrt{n}} \\ &= \frac{1.21}{\sqrt{5}} \\ &= .541.\end{aligned}$$

Hence,

$$\begin{aligned}LCL &= 17.0 - 3(.541) \\ &= 15.4 \text{ ounces} \\ UCL &= 17.0 + 3(.541) \\ &= 18.6 \text{ ounces.}\end{aligned}$$

Example 11.12 To simplify the computation of control limits, special tables have been prepared. These tables are based on the following relationship:

$$3\hat{\sigma}_{\bar{x}} = \frac{3\hat{\sigma}}{\sqrt{n}}.$$

But,

$$\hat{\sigma} = \frac{\bar{R}}{d_2}.$$

Therefore,

$$3\hat{\sigma}_{\bar{x}} = \frac{3\bar{R}}{d_2\sqrt{n}}.$$

However, 3 is a constant as are n and d_2 for samples of a given size. Therefore, a value A_2 may be defined as

$$A_2 = \frac{3}{d_2\sqrt{n}}.$$

Values of A_2 have been computed for different values of n and are shown in Table 11.3.

Since $3\hat{\sigma}_{\bar{x}}$ is equal to $A_2\bar{R}$, the formulas for the control limits may, therefore, be written as

$$LCL = \hat{\mu} - A_2\bar{R} \qquad UCL = \hat{\mu} + A_2\bar{R}.$$

If we apply these formulas to the canning process data we obtain,

$$\begin{aligned} LCL &= 17.0 - (0.577)(2.82) & UCL &= 17.0 + (0.577)(2.82) \\ &= 15.4 \text{ ounces} & &= 18.6 \text{ ounces.} \end{aligned}$$

The results, as expected, are the same as those computed without the use of the A_2 values.

The process average line and the upper and lower control limit lines are drawn on the control chart in Figure 11.4. The values for the 25 sample means are also plotted. Since all the sample means fall within the control limits, we conclude that the process of filling cans is operating in statistical control, and, hence, the control limits, as calculated, may be used to control the process in the future.

We turn now to the control chart of the range. The central line on the chart is equal to $\bar{R}$. As, heretofore, we shall set the control limits at a distance of $3\sigma_R$ from $\bar{R}$. Thus

$$LCL = \bar{R} - 3\sigma_R \quad \text{and} \quad UCL = \bar{R} + 3\sigma_R.$$

It can be shown that

$$\sigma_R = \sigma_w \times \sigma,$$

where σ_w is the standard deviation of the relative range—the sample range divided by the standard deviation of the population. Values of the standard deviation of the relative range have been tabulated. They are constant for a given sample size. Since σ is unknown, we can estimate it from the average sample range, and, hence, we may write

$$\sigma_R = \sigma_w \frac{\bar{R}}{d_2}.$$

We may, therefore, write the control limits as

$$\begin{aligned} LCL &= \bar{R} - \frac{3\sigma_w \bar{R}}{d_2} & UCL &= \bar{R} + \frac{3\sigma_w \bar{R}}{d_2} \\ &= \left(1 - \frac{3\sigma_w}{d_2}\right)\bar{R} & &= \left(1 + \frac{3\sigma_w}{d_2}\right)\bar{R}. \end{aligned}$$

Since σ_w and d_2 are constants for a given sample size and 1 and 3 are constants, the expressions in the parentheses are constants for a given sample size. Thus, we define

$$D_3 = \left(1 - \frac{3\sigma_w}{d_2}\right) \text{ and } D_4 = \left(1 + \frac{3\sigma_w}{d_2}\right).$$

Since, the range can never be less than zero, D_3 is given a value of

zero whenever the above expression for D_3 yields a negative value.

The control limits for the range may now be written as

$$LCL = D_3\bar{R} \quad \text{and} \quad UCL = D_4\bar{R}.$$

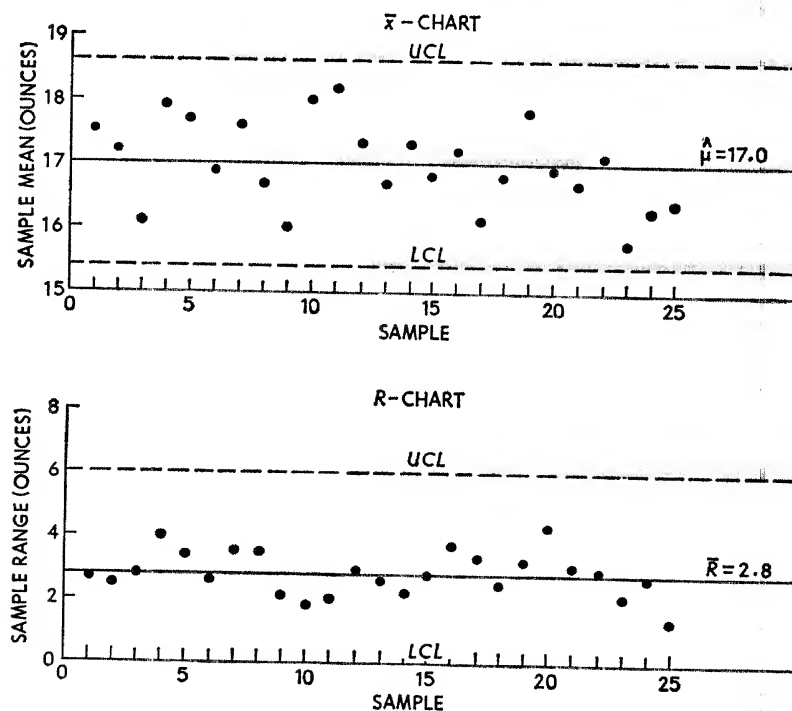
Values of D_3 and D_4 for selected sample sizes are shown in Table 11.3 (p. 346). The control limits of the R -chart for the canning process illustration are

$$\begin{aligned} LCL &= (0)(2.82) & UCL &= (2.11)(2.82) \\ &= 0 & &= 6.0 \text{ ounces.} \end{aligned}$$

The control limit lines are shown in Figure 11.4, as are the values of the 25 sample ranges. The chart indicates that the process is in statistical control, with respect to variability. Since both the $\bar{x}$ -chart and R -chart indicate statistical control, they may be used as guides for taking future action on the process. Note carefully that both charts must indicate control. If either chart or both charts indicate a lack of control, the process must be investigated and brought into control.

Figure 11.4

CONTROL CHARTS FOR MEAN AND RANGE FOR DRAINED WEIGHT OF CONTENTS OF SIZE NO. 303 CANS OF GOLDEN SWEET KERNEL CORN



The decision on whether a p -chart or a combination of $\bar{x}$ - and R -charts is appropriate is not always clear cut. Thus, a company checks on a process of cutting steel rods of a given length. It may check the quality of performance by means of go and not-go gages. That is, if the rod is too short, it will go through one gage. On the other hand, if the rod is too long, it will not go through a second gage. Every rod that is tested by the gages is classified as defective or nondefective, and the p -chart is appropriate.

The rod, however, could be checked by actual measurement of its length—for example, to the nearest tenth of an inch. In such instances a combination of the $\bar{x}$ -chart and R -chart is appropriate.

What considerations determine the choice of the type of control chart in such cases? For one thing, the p -chart is not as sensitive a measure for detecting shifts in the nature of a process as the combination of $\bar{x}$ - and R -charts. Samples of 4 or 5 are usually sufficient for mean and range charts to detect shifts in process level or process variability. A much larger sample is needed for a p -chart.

Example 11.13 The length of a 10-inch rod is checked by go and not-go gages. If the process proportion defective is 1 per cent, a sample of 4 would fail to detect a shift to 10 per cent defective approximately two thirds of the time. That is, about two thirds of the time samples of 4 would have no defective rods even though the process was turning out rods 10 per cent of which were defective.

If π , the proportion defective, were equal to .10, the probability of selecting a nondefective rod would be .90. Hence, the probability of selecting 4 nondefective rods in a sample of 4 would be

$$.90 \times .90 \times .90 \times .90 = .6561.$$

Thus, about two thirds of the time samples of 4 would indicate that there is nothing wrong with the process even though as many as 10 per cent of the rods produced were defective.

While $\bar{x}$ - and R -charts are more sensitive, they usually are more costly to construct than p -charts. More skilled and better trained inspectors are needed to make measurements than are needed to use go and not-go gages. The p -chart may be used alone; the $\bar{x}$ - and R -charts must be used in combination. Furthermore, if more than one quality characteristic must be measured, one p -chart might suffice while separate $\bar{x}$ - and R -charts would be required for each characteristic. Thus, if we were required to control the length of the rod, the diameter of the rod, and its breaking strength, a single p -chart might be used. If either the length, diameter, or breaking strength were not proper, the rod would be called defective. If

$\bar{x}$ - and R -charts were used, three pairs of charts would be needed to control the length, diameter, and breaking strength. The decision on which chart to use in situations where a choice is possible must depend on the balancing of costs as against possible savings.

The $\bar{x}$ - and R -charts have many uses in addition to the control of production processes. Two such illustrations follow.

Example 11.14 A firm desires to control the amount of overtime work in its plant. It uses a combination of $\bar{x}$ - and R -charts for this purpose.

The 5 daily observations of hours of overtime worked constitutes a single sample. Thus, suppose that the number of hours of overtime worked during a week is as follows:

	Mon.	Tues.	Wed.	Thurs.	Fri.
Overtime hours:	40	94	120	80	52

The sample mean ($\bar{x}$) is 77.2 hours, and the sample range (R) is 80 hours.

With these basic data, control charts are constructed in the usual manner with 3-sigma limits. Whenever the average number of hours of overtime, or the range of hours, or both falls outside of control limits, a special investigation of the overtime during that week is undertaken. At times out-of-control points may be explained by rush orders or a breakdown of machinery during regular working hours. At other times it might be found that excessive overtime was caused by poor production scheduling.

Example 11.15 A cost accountant employs $\bar{x}$ - and R -charts to control the amount of waste material in a machine operation. The department has 10 machines. Every day the average weight of scrap for the 10 machines ($\bar{x}$) and the range in weights (R) is determined. These form the basic data for the control charts.

The charts are prepared and used in the usual manner.

11.5 The Control Chart for Number of Defects

There are occasions when neither the combination of the $\bar{x}$ - and R -chart nor the p -chart are satisfactory as bases for decisions on a process.

Example 11.16 A manufacturer of television cabinets desires to control his production process. To do so he counts the number of defects. Defects include scratches, nicks, dents, unpainted portions, poor joints, and so on. The defects can be counted, but they cannot be expressed as a proportion because there is no method of determining the total number of opportunities for defects.

Example 11.17 A company desires to maintain control of its accident rate. The attainment of safety may be looked upon as a proc-

ess. An accident is a defect in the process. Again, we can count the number of accidents, but we do not know the denominator that is needed to express the number of accidents as a proportion.

The basis for decision on a process in situations such as those described in the last two examples is the control chart for the number of defects—usually called the *c*-chart. The central line on the *c*-chart is the average number of defects, C . Thus, if a control chart were to be used to check the process of assembling an airplane wing, C would represent the average number of defects per wing for all wings that might be assembled by the same process. The 3-sigma limits are set at $\pm 3\sigma_c$ from C . The question now is: "How do we estimate C and σ_c ?"

We may say that

$$C = n\pi.$$

That is, C is equal to the probability that a defect will occur multiplied by the number of opportunities for such a defect to occur. In the case of the airplane assembly, we may assume that n , the opportunity for a defect to occur, is large—a wing does encompass a wide area. We can further state that π is very small—otherwise the assembly process would not result in useable wing assemblies. We may make such assumptions even though we do not know the exact value of either n or π . We can, however, estimate their product—namely, C .

For any given wing assembly we can say that the number of defects (c),

$$c = np.$$

The expected value of np is $n\pi$. Hence, the expected value of c is equal to C . Furthermore, the standard deviation of np is

$$\begin{aligned}\sigma_{np} &= \sqrt{n\pi(1-\pi)} \\ &= \sqrt{n\pi - n\pi^2}.\end{aligned}$$

But, when π is small, $n\pi^2$ is very small—very close to zero—and may be disregarded. Hence, we may write

$$\sigma_{np} = \sqrt{n\pi} \quad \text{or} \quad \sqrt{C}$$

We may conclude, therefore, that the sampling distribution of c is approximated by a distribution for which $E(c)$ is equal to C and in addition, for which, σ_c is $\sqrt{C}$. Since the mean and the variance are equal, it is necessary only to estimate C to construct a control chart for the number of defects.

Example 11.18 The number of accidents in a plant varies from week to week. The number of workers employed and the hours worked remain fairly constant from one week to the next. Thus, the exposure to accident remains constant. The firm wants to use a control chart—the *c*-chart—to determine when and when not to take action on its safety program.

Since it would be impractical to wait 25 weeks to obtain 25 points that are usually required to set up a control chart, 13 points are used initially to set up the control chart. The number of accidents is defined as the number of first-aid cases. For the first 13 weeks, these are reported as follows:

<i>Week Ending</i>	<i>Number of Accidents</i>
Jan. 9.....	13
16.....	12
23.....	9
30.....	15
Feb. 6.....	11
13.....	16
20.....	17
27.....	15
Mar. 6.....	17
13.....	15
20.....	16
27.....	14
Apr. 3.....	12
Total.....	182

The average number of accidents per week is estimated from the sample data in the preceding example as follows:

$$\begin{aligned}\hat{C} &= \frac{\text{TOTAL NUMBER OF ACCIDENTS}}{\text{NUMBER OF WEEKS}} \\ &= \frac{182}{13} \\ &= 14.0\end{aligned}$$

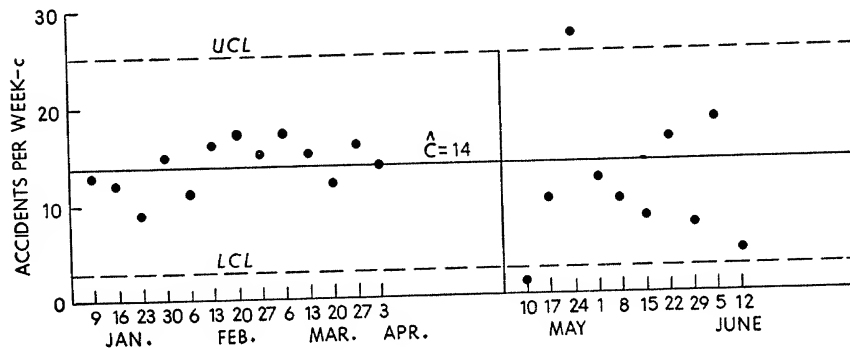
The upper and lower control limits may now be determined as follows:

$$\begin{aligned}LCL &= \hat{C} - 3\hat{\sigma}_c & UCL &= \hat{C} + 3\hat{\sigma}_c \\ &= \hat{C} - 3\sqrt{\hat{C}} & &= \hat{C} + 3\sqrt{\hat{C}} \\ &= 14 - 3\sqrt{14} & &= 14 + 3\sqrt{14} \\ &= 14 - 11.2 & &= 14 + 11.2 \\ &= 2.8, & &= 25.2.\end{aligned}$$

The *c*-chart for the number of accidents is shown in Figure 11.5. You will observe that all points fall within the control limits. In other words, the number of accidents per week vary because of com-

mon causes, and the safety program of the firm requires no special investigation. Furthermore, the firm may project the control chart into the future to check on its safety program.

Figure 11.5



Let us now suppose that the following data are compiled for the ensuing 10 weeks.

<i>Week Ending</i>	<i>Number of Accidents</i>
Apr. 10.....	1
17.....	10
24.....	27
May 1.....	12
8.....	10
15.....	8
22.....	16
29.....	7
June 5.....	18
12.....	4

For the week ending April 10, the control chart in Figure 11.5 indicates that a point was below the lower control limit. The matter was investigated, and it was found that minor accidents were not reported by a newly hired plant physician. The physician was, therefore, instructed as to proper accident reporting procedures.

For the week ending April 27, a point was above the upper control limit. This time an investigation revealed that workers were neglecting to wear safety glasses. The matter was corrected.

11.6 The Risks of Incorrect Action

The control chart, as we have seen, represents a statistical decision rule for determining whether or not to take action on a proc-

ess. Since a control chart makes use of sample data, the following risks of taking incorrect action invariably exist:

- a) The risk of taking action on the process if no trouble exists—the α -risk.
- b) The risk of not taking action on the process when trouble does exist—the β -risk.

A decision rule may be determined by setting the α -risk and β -risk that we can tolerate relative to two possible states of the world and then determining the required sample size and critical values—control limits—of the pertinent statistic. An alternate method of determining a decision rule is to set the sample size and the α -risk and then determine the critical values. It is this latter course that is followed in the construction of control charts.

In practice, as you have seen, 3-sigma control limits are usually used in preference to the direct specification of the α -risk in terms of probabilities. The use of 3-sigma limits rests, to a large extent, on the fact that the sampling distribution of the statistics used are only approximated by known sampling distributions. Thus, the sampling distribution of the mean is distributed normally only if the population from which samples are drawn is normal. In addition, to measure probabilities precisely, the mean of the population and its standard deviation must be known. However, in constructing a control chart the process mean and standard deviation are estimated and, hence, subject to sampling error. Thus, in setting the control limits at $\pm 3\sigma_x$ from $\hat{\mu}$, we cannot say that the probability that sample means will fall outside of control limits is precisely .0027 if the process average is μ . We can, however, say that the probability of a sample mean falling outside of control limits is very small if the process average is μ . The 3-sigma limits set a low probability for the α -risk—its exact value is unknown in most cases.

Similar considerations apply to the other control charts. Thus, the sampling distribution of p is only approximated by the normal curve even for large samples. Similarly, the normal curve offers only an approximation to the sampling distribution of the number of defects. Also, the values for D_3 and D_4 used for the R -chart assume that the population is normal.

It is, therefore, argued that precise probability limits are not needed, since the probabilities are at best approximations. Furthermore, the use of 3-sigma limits as a standard for all control charts makes the technique simpler to use with plant personnel who are untrained in statistics. Finally, about 20 years of experience have

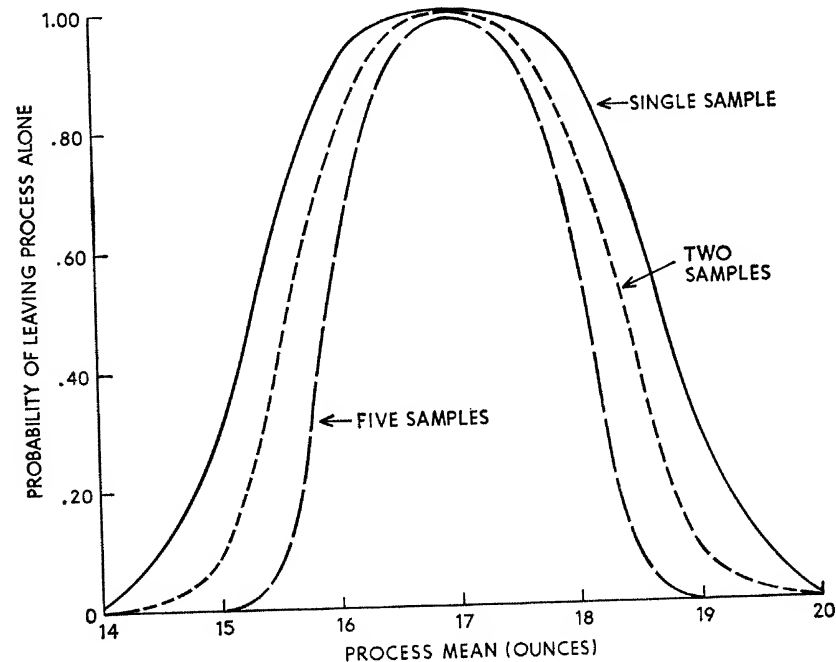
shown the practical value of the 3-sigma control limits. Their use has made for an economic balancing of the risks of possible incorrect decisions.

After the control chart (decision rule) has been determined, it is possible to calculate the probability of the second-type error—leaving the process alone when special causes are present—for all possible values of the decision-parameter. In other words, we can determine the *OC* curve for the control chart.

Example 11.19 Figure 11.6 represents the *OC* curves for the $\bar{x}$ -chart in the process of filling cans of corn in Example 11.11. The estimated process average is 17.0 ounces. The *UCL* is 18.6 ounces, and the *LCL* is 15.4 ounces. The estimated process standard deviation is 1.21 ounces.

Figure 11.6

OPERATING CHARACTERISTIC CURVE FOR CONTROL CHART FOR
SAMPLE MEAN DRAINED WEIGHT OF SIZE NO. 303 CANS
OF GOLDEN SWEET KERNEL CORN



The solid line represents the *OC* curve for a single sample. That is, it shows the probability that a single sample drawn from the process will result in leaving the process alone. The curve indicates that if the process average net weight of cans were to shift 1 ounce to 16.0 ounces (or 18.0 ounces), the probability that the process would be

left alone—the shift would be undetected—is .87. The probability of a shift of 2 ounces remaining undetected by a single sample would be only .23. The probability of not detecting a shift of 3 ounces or more is practically zero. In other words, a shift of 3 ounces or more would be detected practically always with a single sample. In general, therefore, the greater the shift of the process average from its desired level, the more likely that this shift will be detected by a single sample.

The control chart, however, relies on periodic sampling from a process. Successive samples facilitate the detection of sustained shifts in the process average. Thus, the probability of a shift of 1 ounce remaining undetected by a single sample is .87. The probability that such a shift remains undetected in 2 successive samples is $.87 \times .87$, or .76. In general, the probability that any shift remains undetected in m samples may be expressed as $(P_a)^m$, where P_a is the probability of a shift of a given amount in the process average remaining undetected by a single sample. Thus, the probability that a shift of 1 ounce remains undetected after 5 samples is $(.87)^5 = .50$. That is, a shift of 1 ounce will remain undetected .50 times in 100 with 5 successive samples.

In Figure 11.6, *OC* curves have been drawn to show the probabilities of leaving the process alone—a shift in the process remaining undetected—after 2 and after 5 successive samples. The chart indicates that successive samples result in rapid decreases in the probability that shifts in the process average will not be detected. Thus, a sustained shift of 2 ounces will practically always be detected with 5 successive samples.

Since samples are drawn regularly for control charts, large sustained shifts in the process average are almost certain to be detected after a few samples. The probability of detecting small sustained shifts in the process average also increases, but more slowly. Thus, knowledge of the *OC* curve for a control chart is helpful in determining the frequency with which to sample. The more important it is to detect shifts in the process average, the more frequently samples should be drawn.

The *OC* curve of the $\bar{x}$ -chart has been used to illustrate the discussion in this section. The same conclusions, however, are valid for all the control charts considered in this chapter.

11.7 The Samples, Their Size, and Frequency of Selection

Each sample taken from a process represents a subgroup of observations. The successful operation of control chart techniques de-

depends upon the proper grouping of sample observations into subgroups, the size of these subgroups, and their frequency of selection.

Remember that the purpose of a control chart is to detect the presence of assignable causes of variability. To serve this purpose adequately, samples should consist of subgroups of output that are as homogeneous as possible. That is, every observation in a sample should reflect the influences of the same set of causes. Thus, when assignable causes do appear, they will manifest their existence in differences among the samples. If subgroups are chosen so that all samples reflect assignable and unassignable causes, the presence of the assignable causes escapes detection since their effects are submerged. Subgroups, in other words, must be chosen so that the variability among subgroups exceeds the variability within a subgroup.

Thus, it would not be desirable to include in a sample the output of two or more shifts. In this way an unassignable cause in one shift might escape detection on the control chart. Similarly, mixing the product of more than one machine or more than one worker might conceal the assignable cause present in the output of one machine or one worker. Let us use an illustration to explain this point more clearly.

Example 11.20 Two machines are supposed to be cutting tubing into 8-inch lengths. The lengths of the tubing turned out by these machines, to the nearest inch, are given in Table 11.4.

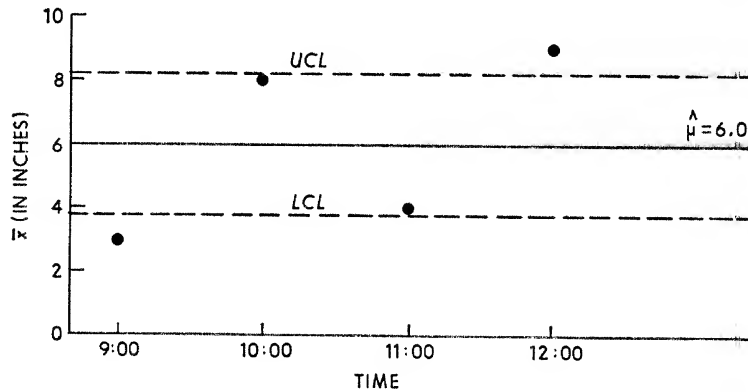
Table 11.4
LENGTH OF TUBING CUT BY MACHINES
(In Inches)

Time	Machine A	Machine B
9:00	3	8
	5	8
	2	8
	2	8
10:00	6	8
	10	8
	9	8
	7	8
11:00	5	8
	3	8
	4	8
	4	8
12:00	10	8
	10	8
	8	8
	8	8

The product of machine B is obviously in control since the product (in defiance of nature) is perfectly uniform. The control chart of the mean based on 4 samples of size 4 for the output of machine A is shown in Figure 11.7.

Figure 11.7

CONTROL CHART FOR $\bar{X}$ FOR OUTPUT OF MACHINE A
SAMPLE SIZE—4

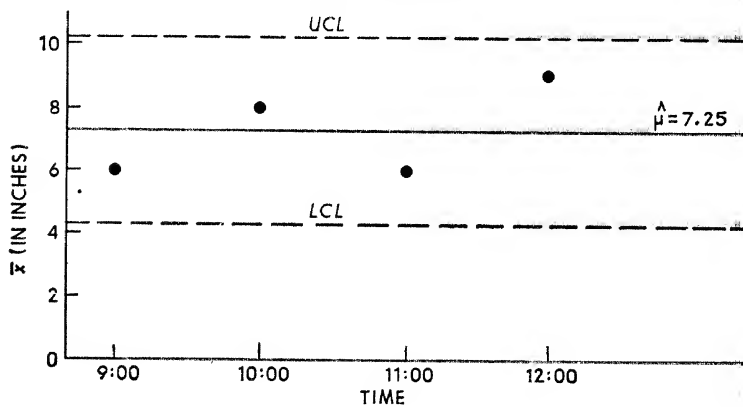


The output of machine A is definitely not in statistical control since 2 of the 4 samples fall outside the control limits.

Instead of taking samples of 4 from each machine separately, let us take 2 items from each machine to make up the sample of 4. Assume that the first 2 items listed in Table 11.4 are selected every hour for each of the machines. The control chart for the combined samples is shown in Figure 11.8. The chart indicates a process in statistical control.

Figure 11.8

CONTROL CHART FOR $\bar{X}$ FOR OUTPUT OF MACHINES A AND B
SAMPLE SIZE—4



In this simplified example, we have shown that by combining in one sample a process operated by random forces and one affected by assignable causes, the presence of the assignable cause remains undetected. The averaging process conceals the changing cause system in A.

It is, therefore, most important in selecting samples that a rational choice of subgroups be made. The choice of the proper subgroup is obviously the engineer's job. However, the statistician must point out its importance for the effective utilization of the control chart.

For the control charts of the mean and range sample sizes of 4 or 5 are used almost universally. One reason follows from our previous discussion. There is greater likelihood that a smaller sample will be more homogeneous. Furthermore, for the same frequency of sampling, samples of 4 or 5 are less costly than larger samples.

On the other hand, larger samples give better protection with respect to the risk of leaving the process alone when assignable causes are at work. Since the standard error decreases with an increase in sample size, the control limits are brought closer together for larger samples. This decreases the range of the statistic over which the process will be left alone. The smaller this range, the smaller the likelihood of leaving the process alone by error. Similarly, the greater the frequency with which samples are taken, the greater the probability the assignable causes will be detected.

The nature of subgroups, the sample size, and frequency of sample selection, require careful consideration to ensure effective operation of control charts. There is as yet no precise solution as to how frequently to sample and as to how large samples should be to obtain maximum protection. The answer to this problem requires a knowledge of the costs connected with the sample size, and frequency of sampling, as well as the costs of undetected process shifts. In addition, knowledge of the probability of such shifts is required.

11.8 Statistical Control and a Satisfactory Product

When we say that a process is in statistical control, we mean that the pattern of its variability is attributable to common causes. When we say that a product is satisfactory, we mean that each unit of the product meets the specification set for it. Statistical control is determined through the use of control limits. These, you remember,

refer to the limits of chance variability for sample means, sample proportions, sample ranges, sample standard deviations, or the number of defects in a sample. Specification or tolerance limits, however, refer to the extent of permissible variability in the *individual items*. If a process is operating in statistical control, it does not follow that all the individual items will be satisfactory, i.e., meet specifications. That is, the variability of the items, although caused by chance factors, might be too great to meet the requirements set for product. Let us consider an illustration:

Example 11.21 Let us assume that the specification for size No. 303 Golden Sweet Whole Kernel Corn is set as 17 ± 1.7 ounces. That is, the drained weight of each can should be 17 ounces, with an allowance of 10 per cent deviation each way. The lower specification is set to protect consumers. The cannery, on the other hand, does not want to give the consumer more than he is paying for and, hence, sets an upper specification. Any can, therefore, is considered acceptable if its drained weight varies from 15.4 to 18.6 ounces.

The control charts in Figure 11.4 indicated that the process of filling cans was in control with respect to the process average and the process variability. Nevertheless, if you refer back to Table 11.2 you will observe that 26 of the 125 sampled cans were not acceptable—12 cans weighed less than 15.4 ounces, and 14 cans weighed more than 18.6 ounces. We, thus, see that a process in control can produce items that are not satisfactory—that do not meet the specifications.

Specification or tolerance limits are determined so that the product will be serviceable. Thus, for example, if a cylinder is to fit snugly over a tube, the specifications for the inner diameter of the cylinder must be set so that it covers the tube but does not fit too loosely. Some variability must be allowed for in the specifications, for, as we know, all production processes are subject to variability. What constitutes a serviceable product is a problem not for the statistician but for the engineer. However, the statistician can aid, as we shall see, in helping to determine specification limits that are realistic.

What are the alternatives open to a firm when the process variability is attributable to common causes (operating in statistical control) but when too large a portion of the individual items do not fall within the specification limits? An obvious answer would be to change the process. Since the process is in statistical control, any tampering with it might only make matters worse. A change in process, however, is usually a costly undertaking. Such changes might

include the purchase of new machinery, the design of new tools, and the retraining of workers.

If the firm does not change the process, then it might resort to 100 per cent inspection to separate the good product from the bad. This also entails costs and is not possible when a destructive test must be used to distinguish between good and bad product. Management must decide, on the basis of costs, which of the alternatives it is to follow.

At times, a third alternative is available. It has been found that producers often set specifications too tightly for the purpose the product is to serve. A review of such specifications has disclosed that the limits might be relaxed.

It is, therefore, preferable to determine process capabilities before specification limits are set. Then, before production begins it is possible either to adjust (1) the specifications to the capabilities of the process; (2) to consider the possibility of changing the process; or (3) to use 100 per cent inspection and to scrap or rework all defective product.

The control chart can be used as an aid in determining process capability. Before a process gets into production a pilot run can be made. The pilot run accomplishes two things. In the first place, it enables us to determine whether the process is in control. If it is in control, it enables us to estimate the process capability—a measure of the variability in the individual items.

Example 11.22 A manufacturer is to produce a temperature control device. Forty devices are produced and set so that the switch is to operate at 68.0 degrees. The temperature at which the switch actually operates is recorded for the 40 devices. The devices are grouped into 10 groups of 4; $\bar{\mu}$ is found to be 68 degrees, and $\bar{R}$ is found to be 0.9 degrees. The control charts for the sample mean and the sample range are plotted, and the process is found to be in statistical control. The question we wish to answer is: "What is the process capability?" You again observe that we could not predict process capability if the process were not in statistical control.

We determine the variability in the individual devices (process capability) as follows: With $\bar{R}$ and the sample size ($n = 4$) we can estimate the process standard deviation ($\hat{\sigma}$).

$$\begin{aligned}\hat{\sigma} &= \frac{\bar{R}}{d_2} \\ &= \frac{0.9}{2.059} \\ &= .44\end{aligned}$$

If we now assume that the distribution of temperatures at which the switch operates is normal, we can determine the process capability. We know that all but approximately .0027 of the temperatures will fall within the range of $\mu - 3\hat{\sigma}$ to $\mu + 3\hat{\sigma}$.

$$\begin{aligned}\mu - 3\hat{\sigma} &= 68 - 3(.44) = 66.7^\circ, \\ \mu + 3\hat{\sigma} &= 68 + 3(.44) = 69.3^\circ.\end{aligned}$$

The manufacturer wanted to set his specifications at $68^\circ \pm 0.5^\circ$. Do you think that the process is capable of meeting these specifications?

You should observe that these computations do not tell you what the specification limits should be. They do tell you what type of specification limits a process could meet.

11.9 You Should Now Know That

Variability is a characteristic of all business processes.

The causes of process variability may be divided into two broad categories: (a) common (unassignable, uncontrollable) causes, and (b) special (assignable, controllable) causes.

Common causes are inherent in the nature of a process and are usually difficult to isolate and to correct.

A process whose variability can be accounted for by common causes is said to be in statistical control.

Special causes usually are identifiable and can be corrected.

A process whose variability can be accounted for by special causes is said to be out of statistical control.

A control chart is a statistical decision rule presented in the form of a graph to distinguish between common and special causes. In this way, it serves as a basis for deciding whether or not to look for special causes.

A control chart may be designed only for processes that have been brought under statistical control.

The most commonly used control charts are: the p -chart, the $\bar{x}$ -chart, the R -chart, and the c -chart. The $\bar{x}$ - and R -charts are usually used in conjunction with each other.

Since control charts are based on sample information, two types of incorrect action are possible: (a) taking action on the process when no trouble exists, and (b) neglecting to take action on a process when trouble does exist.

The probability of taking action on a process when no trouble

exists (α -risk) often is set implicitly by the use of 3-sigma limits for all types of control charts.

The operating characteristic curve of a control chart presents the probability of not looking for special causes for various shifts in the nature of the process.

The samples drawn for use with control charts should represent homogeneous subgroups of the population. The sample size and the frequency with which samples are drawn are also important considerations in assuring efficient use of control charts.

Statistical control does not imply a satisfactory product. If a process in control does not result in a satisfactory product, a change in the process or 100 per cent inspection generally is called for. At times, however, specifications might be revised.

The control chart, by determining process capability, may be used to predict the ability of a process to produce a product to meet given specifications. However, the control chart does not determine specifications.

11.10 You Should Now Be Able to Solve

1. What type of control chart(s) would you recommend in each of the following situations? Explain the reasons for your choice.

- a) A company wants to control the error rate in its filing system.
- b) A company wants to control the number of imperfections in a bolt of cloth.
- c) A company wants to control the breaking strength of glass containers.

2. A firm manufactures bolts. It decides to use a control chart to control the diameters of various size bolts.

- a) Could the firm use $\bar{x}$ - and R -charts? Explain.
- b) Could the firm use a p -chart? Explain.
- c) What factors would the firm have to consider in deciding between a combination of $\bar{x}$ - and R -charts and a p -chart?

3. A company decides to use a p -chart to control its process of invoice preparation. Random samples of 100 invoices are to be drawn twice daily, and the proportion of incorrect invoices in each sample is to be recorded. An invoice is to be considered as incorrect if it contains an omission, incorrect name of account, incorrect account number, erasure, strike-over, transposition of figures, incorrect quantity, incorrect unit price, incorrect extension, or incorrect total. To construct the initial control chart 25 samples are drawn, and the results are found to be as follows:

Date	Sample Number	Sample Size	Number of Incorrect Invoices
March 2	1	100	3
	2	100	5
3	1	100	2
	2	100	3
4	1	100	2
	2	100	6
5	1	100	0
	2	100	4
6	1	100	11
	2	100	12
9	1	100	5
	2	100	3
10	1	100	4
	2	100	1
11	1	100	11
	2	100	1
12	1	100	6
	2	100	4
13	1	100	2
	2	100	3
16	1	100	2
	2	100	3
17	1	100	2
	2	100	12
18	1	100	3

- Construct the initial p -chart.
- Is the process of invoice preparation under statistical control? Explain.
- What must be done before a control chart may be prepared for future use?
- Set up the p -chart for future use.

4. Draw an OC curve for the p -chart constructed in Problem 3 for a single sample. For 2 samples. For 5 samples. (Assume that the sampling distribution of p is normal.)

5. A company wishes to check on the reliability of laboratory inspection results. A standard sample of 4 items is submitted at random intervals for inspection (unknown to the laboratory techni-

cian). The technician determines the per cent of metal by weight in the 4 units of the compound submitted to him for analysis. The results of the first 24 submissions of the same standard sample are shown below:⁴

Standard Sample Number	Per Cent of Metal by Weight			
	Item 1	Item 2	Item 3	Item 4
1.....	10.28	10.30	10.25	10.25
2.....	10.26	10.26	10.19	10.30
3.....	10.27	10.28	10.25	10.26
4.....	10.28	10.30	10.28	10.30
5.....	10.15	10.28	10.43	10.20
6.....	10.13	10.23	10.29	10.37
7.....	10.27	10.39	10.28	10.29
8.....	10.20	10.29	10.12	10.25
9.....	10.25	10.33	10.20	10.24
10.....	10.25	10.31	10.20	10.21
11.....	10.16	10.20	10.34	9.97
12.....	9.88	10.13	9.97	9.88
13.....	10.13	10.25	10.31	10.25
14.....	10.39	10.25	10.31	10.27
15.....	10.36	10.24	10.31	10.26
16.....	10.26	10.29	10.32	10.20
17.....	10.24	10.27	10.30	10.34
18.....	10.39	10.37	10.21	10.24
19.....	10.31	10.28	10.38	10.23
20.....	10.33	10.26	10.18	10.21
21.....	10.21	10.14	10.21	10.24
22.....	10.14	10.10	10.22	10.25
23.....	10.24	10.12	10.13	10.10
24.....	10.13	10.18	10.22	10.19

- a) Why are both an $\bar{x}$ -chart and an R -chart needed in the present situation?
- b) Set up an $\bar{x}$ -chart on the basis of the above data.
- c) Does the $\bar{x}$ -chart indicate that the process is in statistical control? Explain.
- d) What must be done before an $\bar{x}$ -chart may be prepared for future use?
- e) Set up an R -chart on the basis of the above data.
- f) Does the R -chart indicate that the process is in statistical control? Explain.
- g) How does your interpretation of a point out of control on an $\bar{x}$ -chart differ from your interpretation of a point out of control on an R -chart?
- h) Revise the $\bar{x}$ -chart and R -chart for future use.

⁴Based on Frank W. Kroll, Jr. "Effective Quality Control Program for the Industrial Control Laboratory," *Industrial Quality Control*, Vol. XIV, No. 6 (December, 1957), pp. 9-13.

6. A company manufactures paper that is intended to be impervious to oils. It conducts pinhole tests at regular intervals. A sheet of paper, 17 by 22 inches in size, is taken from production, and colored ink is applied to one side of the sheet. Each ink spot which appears on the other side of the sheet within 5 minutes is counted as a defect (a pinhole).

The results for the first 25 sheets sampled are shown below:

Sheet Number	Number of Pinholes	Sheet Number	Number of Pinholes
1.....	9	14.....	5
2.....	8	15.....	4
3.....	9	16.....	12
4.....	12	17.....	6
5.....	8	18.....	21
6.....	5	19.....	9
7.....	7	20.....	6
8.....	6	21.....	9
9.....	5	22.....	20
10.....	9	23.....	19
11.....	6	24.....	8
12.....	8	25.....	7
13.....	7		

- Construct a *c*-chart on the basis of the above data.
- Is the process in statistical control? Explain.
- How does the control chart help the firm bring its process under control?
- Construct a *c*-chart for future use.

7. A process for the manufacture of steel rods has recently been perfected and is operating in statistical control relative to the breaking strength of the rods. The average breaking strength (process average) is 600 pounds per square inch (p.s.i.), and the process standard deviation is 25 p.s.i.

In spite of the presence of statistical control, complaints have been received from customers concerning the great number of rods that have been breaking.

What might be wrong? What might be done to remedy the situation? In your answer emphasize what statistics can do and cannot do as an aid in coping with the problem.

8. A company manufactures dial knobs. The specification for the inner diameter of the knob is $.50 \pm .03$ inches.

Control charts for the mean and range are set up to control the dimension of the inner diameter. The estimated process average,

based on 25 samples of size 5, is .506 inches, and $\bar{R}$ is .052 inches. The process is found to be in statistical control.

Is the process capable of producing knobs whose inner diameters meet the required specifications? Explain carefully.

11.11 You Will Also Find That

The construction of control charts and their use are discussed very clearly in:

A. S. T. M. *Manual on Quality Control of Materials*. Philadelphia: American Society for Testing Materials, 1951.

American Standard, Z1.3—1958, *Control Chart Method for Controlling Quality during Production*. New York: American Standards Association, Inc., 1958.

GRANT, EUGENE L. *Statistical Quality Control*, chaps. i, ii, iv, v, vii, x, and xi. 2d ed. New York: McGraw-Hill Book Co., Inc., 1952.

For further discussions you may consult:

COWDEN, DUDLEY, J. *Statistical Methods in Quality Control*, chaps. xvi–xix. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1957.

DUNCAN, ACHESON J. *Quality Control and Industrial Statistics*, chaps. xviii–xxii. Rev. ed. Homewood, Ill.: Richard D. Irwin, Inc., 1959.

For varied applications of control chart techniques you should consult *Industrial Quality Control*, a monthly publication of the American Society for Quality Control. The transactions of the Society's annual conventions are also replete with illustrative materials. The articles in these publications are, as a rule, nontechnical.

Comparative Experiments and Their Statistical Design

12.1 The Nature of Experimentation

In this chapter we shall consider some of the statistical methods and theory that are required to reach decisions on the basis of comparative experiments.

An *experiment* may be regarded as a special type of statistical study in the sense that it involves collecting information as a basis for action. In the statistical studies that we have previously considered, observations usually are available and would be available whether or not a statistical study was conducted. In contrast, an experiment ordinarily requires that observations be created—that is, elementary units must be subjected to certain stimuli or influences and their reactions observed. Hence, an experiment usually involves observations on a situation created by the study itself.

To point up this distinction between experiments and other types of statistical studies, contrast a study to estimate the average rent paid for a group of dwelling units with a study to measure the efficacy of a new drug. In the first situation, the observations on rent would be available whether or not a study were conducted. However, in the second, elementary units (such as guinea pigs) must be subjected to the drug in order to determine its influence and, hence, its efficacy.

Many problems requiring experimentation can be decided through the medium of elementary statistical theory such as we discussed in Chapters 7 and 8. Thus, a decision problem on whether or not to market a synthetic plastic might require an experiment to measure its average breaking strength. This situation, however,

simply presents a testing problem for which μ is the decision-parameter. Similarly, the estimation of the percentage of ammonia in a gas may require some experimentation, but the problem simply presents an estimation problem for which π is the decision-parameter.

However, some experiments present the need for more advanced theory and methods. In particular, *comparative experiments* require more complex statistical designs. A comparative experiment involves experimentation as a basis for comparing properties of two or more populations. For example, consider the experiment conducted a few years ago to decide on whether or not to approve the general use of Salk vaccine as a preventive of infantile paralysis. In that instance, one group of children (elementary units) was given Salk vaccine, while another group was not given the vaccine. The problem required that one population (the effects of the vaccine) be compared with another population (the effects of no vaccine). Of course, only a sample from both populations could be observed—the differences shown by the samples of effects served as a basis for decision.

In this chapter, then, we shall consider some aspects of the statistical design of comparative experiments, with particular emphasis on applications in business.

Some persons may associate experiments with chemistry and test tubes, but actually experimentation covers a far wider range of application both in and out of the business world. Thus, pharmaceutical, chemical, and oil companies have a natural interest in laboratory experiments. But, they and many other types of companies also have an interest in such things as marketing and advertising experiments and in industrial experiments where the laboratory may be the factory or the consumer's home or the market place. Why? Important decisions must be based on the results of such experiments. Let us consider some illustrations of business decision problems which require comparative experiments.

Example 12.1 A company develops a new product—a coated magnetic wire. Two methods of coating the wire (that is, two processes) are available for use in production. Since both methods are equally costly, the main criterion for a decision is an answer to the question: "Which method will result in the more uniform product?" An experiment to compare the results of applying method A and of applying method B would provide a basis for the answer to this question and, hence, a basis for decision.

Example 12.2 An advertising agency must decide which of several promotional displays is more effective. As a basis for deciding

which display is more effective and, hence, which should be used, an experiment is to be conducted. Samples of consumers are to be shown the displays; the reactions of the consumers measured through interviews; and the choice of a display made on the basis of these observations.

Example 12.3 A personnel director is responsible for developing a new employee training program. The program is to include some technical instruction, and two methods of presenting the material are available. The director thus faces a decision problem—the choice of a method of instruction. One decision procedure is suggested—use each method of instruction on a sample of employees and, on the basis of the results of this comparative experiment, determine which method of instruction is the more effective.

The statistical design of an experiment is a plan for collecting observations. The proper statistical design of comparative experiments is important in order to control, within economic limits, the risks of incorrect decisions. A good part of this chapter, therefore, will deal with the consideration of alternative designs.

Throughout this text we have emphasized the importance of a priori planning. In short, one must always remember that statistical problems do not begin only after the data are collected. Unfortunately, one sometimes finds that an experimenter collects data and then worries about their analysis. However, not even an expert statistician would be clever enough to analyze data collected haphazardly. It is important, therefore, that the same type of a priori planning stressed throughout our study of statistics be applied in the design of experiments. As the eminent British statistician, Sir Ronald Fisher, points out:

In considering the appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity, what interpretation shall be placed upon each one of them.¹

It is most appropriate to quote Fisher on this topic because his book, *The Design of Experiments*, which was first published in 1935, has become the basis for the development of the subject. Originally applied to the field of agriculture, the principles of experimental design in recent years have been extended for use in the natural sciences, in the social sciences, and in business, commerce, and industry.

¹ Ronald A. Fisher, *The Design of Experiments* (5th ed.; Edinburgh, London: Oliver & Boyd, 1949), p. 12.

12.2 Decisions Involving the Comparison of Two Population Means

A common type of decision problem calls for the comparison of two populations because the consequences of taking an action depend upon the difference between the means of the populations.

Let us denote such populations as A and B, their arithmetic means as μ_A and μ_B , and their standard deviations as σ_A and σ_B , respectively. The difference between the two population means may be called D . Hence,

$$D = \mu_A - \mu_B.$$

We shall consider the type of problem for which the consequences of taking an action depend upon the true, but ordinarily unknown, value of D . Hence, we may say that D describes states of the world, and that the true value of D describes the true state of the world. We assume, then, that if this true difference were known, action could be taken. Let's look at an illustration for future, as well as immediate, reference.

Example 12.4 A manufacturer of fishing equipment must decide on the use of one of two methods of waterproofing fishlines. Thus, the alternatives are: (1) use method A, and (2) use method B. The possible incorrect decisions are: (1) to use method A when actually B should be used, and (2) to use B when actually A should be used.

A good waterproofing method is one which does not lower the breaking strength of the line after its application. Of course, the effects of any method will vary somewhat from fishline to fishline. Hence, two populations must be compared. One population represents the breaking strengths of fishlines treated by method A; the other represents the breaking strengths of fishlines treated by method B. Since this problem, like almost all experiments, is analytical because it calls for action on a process, both populations are infinite.

It is believed that the variability in breaking strengths resulting from the use of method A would be about the same as the variability in breaking strengths resulting from the use of method B. (That is, it is believed that σ_A equals σ_B .) However, there is uncertainty about which method results in the higher average breaking strength. That is, there is uncertainty about the value of

$$D = \mu_A - \mu_B.$$

Some information about this value is necessary in order to make a rational decision. Hence, an experiment is to be conducted to provide sample information about this difference.

Before we attempt to formulate a statistical decision rule for this problem, let us consider the general nature of such an experiment. A sample of the effects of method A and a sample of the effects of

method B are to be observed. The average breaking strength of the fishlines treated by each method then may be computed. Let us denote the possible results of this experiment as follows:

	Sample A	Sample B
Sample size.....	n_A	n_B
Sample mean.....	$\bar{x}_A$	$\bar{x}_B$
Sample variance.....	s_A^2	s_B^2

The pertinent statistic is the difference between the two sample means, namely,

$$d = \bar{x}_A - \bar{x}_B.$$

This statistic provides partial information about the true state of the world—about the true but unknown value of D . The choice of the method of treating fishlines may be related to possible values of d . However, this involves some risks of incorrect action because d represents only sample information. To measure these risks we therefore shall consider its sampling distribution.

First consider this general idea: the sampling distribution of the difference in sample means, for sufficiently large samples, will be approximately normally distributed. In addition, remember that the mean of this distribution, $E(d)$, is

$$E(d) = D.$$

The standard deviation of the statistic d may be denoted as σ_d . The formula for this standard error can be derived by the application of a fundamental theorem of mathematical statistics: The variance of the difference of two independent quantities (such as $\bar{x}_A$ and $\bar{x}_B$) is equal to the sum of the variances of these quantities. Hence,

$$\sigma_d^2 = \sigma_{\bar{x}_A}^2 + \sigma_{\bar{x}_B}^2.$$

We know that the variance of a sample mean for a simple random sample from an infinite population is σ^2/n . Therefore,

$$\sigma_d^2 = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}$$

and

$$\sigma_d = \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}.$$

In many experiments it is reasonable to assume that σ_A equals σ_B . If so, we may denote their common value by σ and write

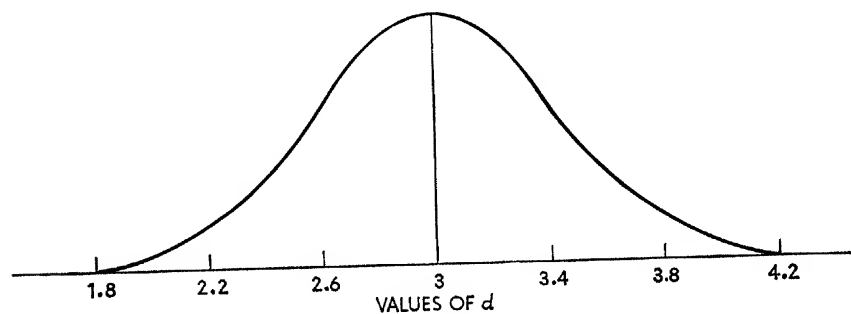
$$\sigma_d = \sigma \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}.$$

This assumption of equal variances will be used in the fishline example that we are considering. Therefore, the last relationship given for σ_d may be used.

Example 12.5 Suppose that we consider a numerical example of the sampling distribution of d in the fishline problem. Let us assume that 50 pieces of fishline are to be treated by method A and that another 50 pieces are to be treated by method B. Their breaking strengths are then to be determined. The difference between the two sample means should be approximately normally distributed with $E(d) = D$ and

$$\sigma_d = \sigma \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}.$$

Hence, if the true difference between μ_A and μ_B were 3 pounds and if the standard deviation of each population were 2 pounds, the sampling distribution of d would appear as follows:



Its mean would be

$$E(d) = D = 3,$$

while its standard deviation would be

$$\begin{aligned} \sigma_d &= \sigma \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \\ &= 2 \sqrt{\frac{1}{50} + \frac{1}{50}} \\ &= 2 \cdot \frac{1}{5} \\ &= .40. \end{aligned}$$

Let us now apply the foregoing theory to the development of a statistical decision rule for the fishline problem. Such a rule must provide the answers to the following four questions:

1. How should the sample be selected?
2. How should the sample be allocated? That is, how many lengths of fishline should be treated by method A and how many by method B?

3. What sample size is required?
4. For what values of d should method A be adopted and for what values of d should method B be adopted? In short, what is the critical value, d^* ?

The answer to the first question is the easiest. The sample, whatever its size, should be random. That is, the lengths of fishline must be chosen and then apportioned between the two methods by the use of random numbers. Some form of randomization is a requirement of every statistical study. Without randomization there is no basis for the application of statistical theory.

The second question calls for an answer to the question of allocating the total sample between the two methods. A general answer to this question for comparative experiments is as follows: If the cost of applying each method is equal and if the two population standard deviations are equal, assign half of the total sample size to each treatment. This allocation minimizes the standard error of d and, hence, minimizes the risks of relying on sample information. To illustrate this thought, reconsider the formula for the standard error of d when $\sigma_A = \sigma_B$,

$$\sigma_d = \sigma \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}.$$

The value of σ_d obviously depends upon the factor

$$\sqrt{\frac{1}{n_A} + \frac{1}{n_B}},$$

and σ_d is minimized when this factor has its lowest value. This occurs when $n_A = n_B$.

Example 12.6 Assume that 100 lengths of fishline are to be water-proofed. Let us consider different values for n_A and n_B and compute the corresponding values of

$$\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}.$$

n_A	n_B	$\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$
10	90	.333
20	80	.250
30	70	.218
40	60	.204
50	50	.200
60	40	.204
70	30	.218
80	20	.250
90	10	.333

We thus observe that $n_A = n_B = 50$ results in the minimum value of

$$\sqrt{\frac{1}{n_A} + \frac{1}{n_B}},$$

and, hence, the minimum value of σ_d .

The way in which we answer the third and fourth questions depends on the objectives of the decision maker. Thus, if the costs of applying the treatments were high, one might wish to set the total sample size, n , directly. Then, the risk of one type of incorrect decision could be set in order to find d^* . Alternately, one might wish to control the risks of both types of incorrect decisions. In this case n and d^* can both be found from the appropriate equations (as in Chapter 8).

Here, let us assume that n is set directly. Thus, in the fishline problem suppose that it is decided to select a total sample of 100—50 lines to be treated by each method. To find the critical value, one risk of an incorrect decision must first be set. That is, either of the following must be set:

P (USING METHOD A IF METHOD B BETTER)

OR

P (USING METHOD B IF METHOD A BETTER)

Moreover, one must describe what is meant by “method B better” or “method A better” in terms of a value of D . Thus, suppose that the company feels that method A would be more economical if the average breaking strength resulting from its use exceeds the average resulting from the use of method B by 1 pound (i.e., if $D = 1$). Further, suppose that they are willing to take only a .01 risk of using B if $D = 1$.

Consider, now, possible values of d . We wish to find a critical value, d^* . For values of d equal to or greater than d^* , method A will be used, while for values of d less than d^* , method B will be used. We have a one-tailed decision rule of the form:

$$\begin{array}{c|c} \text{USE METHOD B} & \text{USE METHOD A} \\ \hline d^* \end{array}$$

VALUES OF d

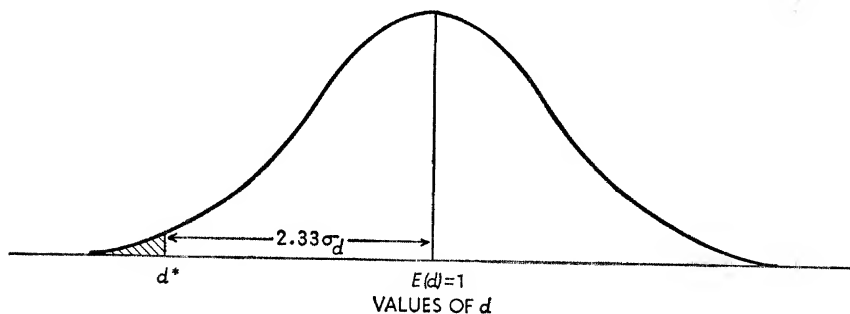
Next let us consider the sampling distribution of d if $D = 1$, keeping in mind the objectives: the sample size,

$$n = 100, \quad n_A = n_B = 50,$$

and the specified risk,

$$P \text{ (USING METHOD B IF } D = 1) = .01.$$

If D equals 1, the sampling distribution of d would be normal with its mean, $E(d)$ equal to 1. Furthermore, d^* must cut off the lower tail of the sampling distribution of d so that only a .01 area is to the left of d^* . Graphically, the situation is as follows:



Because of the tail area of .01, z equals 2.33. Hence, we have

$$\begin{aligned} d^* &= E(d) - 2.33\sigma_d \\ &= 1 - 2.33\sigma_d. \end{aligned}$$

Unfortunately, we do not know the value of σ_d so we cannot immediately compute the critical value. However, once the sample is drawn, σ_d can be estimated from the sample and the value of d^* calculated. Thus, after a sample is drawn, an estimate of σ_d^2 can be computed from the following formula:

$$\hat{\sigma}_d^2 = \frac{n_A s_A^2 + n_B s_B^2}{n_A + n_B - 2} \left(\frac{1}{n_A} + \frac{1}{n_B} \right),$$

where, as noted, s_A^2 and s_B^2 are the two sample variances. Of course, an estimate of σ_d is provided by the square root of the above.

Example 12.7 The experiment to provide a basis for deciding on a method of waterproofing is conducted. The decision rule is: Treat a random sample of 50 pieces of fishline by each method. Observe the average breaking strength in each sample and calculate the difference,

$$d = \bar{x}_A - \bar{x}_B.$$

If d is less than d^* , use method B; otherwise use method A. The critical value determined from the sample will be

$$d^* = 1 - 2.33\hat{\sigma}_d.$$

The results are summarized below:

Method A	Method B
$n_A = 50$	$n_B = 50$
$\bar{x}_A = 15.9$	$\bar{x}_B = 15.2$
$s_A^2 = .115$	$s_B^2 = .128$

Thus, we observe that

$$\begin{aligned} d &= 15.9 - 15.2 \\ &= .7 \text{ pounds.} \end{aligned}$$

In order to translate this result into a course of action, we must compute the value of d^* ,

$$d^* = 1 - 2.33\sigma_d.$$

Thus, first we estimate σ_d :

$$\begin{aligned} \hat{\sigma}_d^2 &= \frac{n_A s_A^2 + n_B s_B^2}{n_A + n_B - 2} \left(\frac{1}{n_A} + \frac{1}{n_B} \right) \\ &= \frac{50(.115) + 50(.128)}{98} \left(\frac{1}{50} + \frac{1}{50} \right) \\ &= .00496 \\ \hat{\sigma}_d &= \sqrt{.00496} \\ &= .070. \end{aligned}$$

Next we compute d^* :

$$\begin{aligned} d^* &= 1 - 2.33 (.070) \\ &= 1 - .16 \\ &= .84 \text{ pounds.} \end{aligned}$$

Lastly we observe that d equals .7 pounds and that this is less than the critical value, $d^* = .84$ pounds. Hence, the decision is to use method B.

Let us summarize our discussion in this section. In general, one way of formulating a statistical decision rule when D is the appropriate decision-parameter is as follows: Decide on the sample size directly. Focus on one value of D (D_0), and specify a tolerable risk of acting incorrectly if D were true. The critical value is then

$$d^* = D_0 + z\sigma_d,$$

where z is the normal deviate required to yield the specified risk, and where σ_d may be estimated from the sample.

Three concluding remarks are in order. First of all, the procedure sketched in this section is applicable only under the assumption that $\sigma_A = \sigma_B$. If this assumption should not be made, the above procedure must be modified. Secondly, the procedure also rests on the assumption that the sampling distribution of d is normal.² This assumption is reasonable if the combined sample size is sufficiently large—as a general rule, if n is at least 25 or 30. If n is small, the sampling distribution of the pertinent statistic ordinarily cannot be assumed normal and special techniques must be incorporated into the statistical decision procedure.

²Strictly speaking, the assumption is that $d/\hat{\sigma}_d$ is normally distributed because we estimate σ_d from the sample.

Lastly, remember that the methods sketched require that n be set directly and that only the risk of an incorrect decision for one possible state of the world (one value of D) be controlled. Therefore, usually it is important to look at the operating characteristic curve of the rule adopted. This OC curve would show risks for other values of D and indicate whether the rule should be modified. While, at times, special problems may arise in the construction of an OC curve for this type of testing procedure, the general method of calculation is similar to that illustrated in Chapter 8.

12.3 Decisions Involving the Comparison of Two Population Proportions

Sometimes comparative experiments are necessary to provide partial information, not on the difference between two population means but on the difference between two population proportions.

Example 12.8 The industrial engineer of a company must design a punch-press operation to stamp out inlays from sheet metal. He considers two alternative methods which differ only in the way the metal is oiled. Method B is slightly more time consuming than method A. However, the engineer is uncertain about the proportion of defectives to be expected from each method. He, therefore, wishes to conduct a comparative experiment to provide a basis for decision.

Let us denote the two population proportions of interest in this type of problem by π_A and π_B . Thus, in the punch-press problem, π_A may be defined as the proportion of defectives that would be produced by method A if that operation were repeated indefinitely under similar conditions. π_B is defined similarly for method B. Furthermore, let

$$D = \pi_A - \pi_B.$$

In Example 12.8, D describes states of the world for the decision problem relating to the two methods of operating the punch press. Thus, we shall consider

$$d = p_A - p_B$$

as the pertinent statistic to be derived from an experiment comparing the two methods. The statistics, p_A and p_B , are the proportions of defectives in the two random samples taken from the two methods of operation.

Hence, consider the sampling distribution of d . For sufficiently large samples, the distribution will be approximately normal with mean equal to

$$E(d) = \pi_A - \pi_B = D$$

and variance equal to

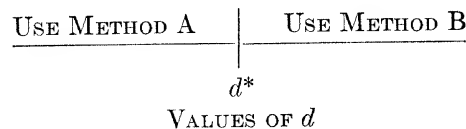
$$\sigma_d^2 = \frac{\pi_A(1 - \pi_A)}{n_A} + \frac{\pi_B(1 - \pi_B)}{n_B}.$$

The latter result follows from the same theory applied to derive the variance of $(\bar{x}_A - \bar{x}_B)$ in the last section.

Now, let us develop a statistical decision rule for the type of problem under consideration. As in the last section, we shall treat the case for which sample size is determined directly and one risk of an incorrect decision is specified.

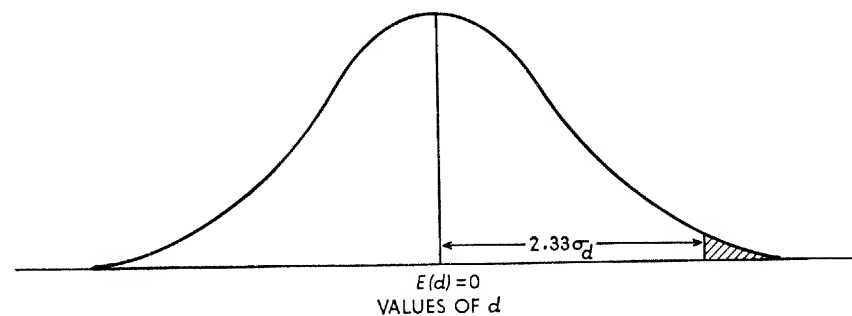
Assume that the engineer in Example 12.8 wishes to spend only about one hour of worker and machine time on the experiment. He estimates that in that time he can get a sample of 5,000 inlays from method A and a sample of 4,800 inlays from method B. Workers are to be instructed to punch that many inlays. (Note that here the cost, in time, of sampling is not equal and a larger sample from the cheaper or faster method is called for.) The engineer feels that because method B is the slower, if $D = 0$ he wishes only a .01 risk of adopting method B.

We must now find a critical value, d^* , which divides all possible results of the experiment—all possible values of d —into two groups each of which leads to a different action:



Thus, values of d less than d^* call for the use of method A, while values of d equal to or greater than d^* dictate the use of method B.

The engineer indirectly has specified this critical value by indicating that if D equaled 0, the tolerable risk of using method B should be only .01. If D is 0, the sampling distribution of d appears as follows:



Its mean, of course, is $E(d) = 0$. Its variance is

$$\sigma_d^2 = \frac{\pi_A(1 - \pi_A)}{n_A} + \frac{\pi_B(1 - \pi_B)}{n_B}.$$

If $D = 0$, then π_A is equal to π_B , and we may write

$$\sigma_d^2 = \pi(1 - \pi)\left(\frac{1}{n_A} + \frac{1}{n_B}\right).$$

The critical value, d^* , as can be seen, is

$$\begin{aligned} d^* &= 0 + 2.33\sigma_d \\ &= 2.33\sigma_d. \end{aligned}$$

This cannot be immediately computed because σ_d depends upon the unknown value of π ($= \pi_A = \pi_B$ by assumption). However, this value can be estimated from the sample, and d^* can be computed at that time. An estimate of π from the samples is provided by the proportion of defectives in both samples, p . Hence, we may use

$$\hat{\sigma}_d^2 = p(1 - p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)$$

as an estimator for σ_d .

Example 12.9 Let us consider the application of the decision rule developed in this section. Suppose that the engineer has 5,000 inlays stamped by method A and 4,800 inlays stamped by method B and that the results are as follows:

Method	Sample Size	Number of Defectives	Proportion Defective
A.....	5,000	255	.0510
B.....	4,800	216	.0450
Total.....	9,800	471	.0481

We now note that

$$\begin{aligned} d &= p_A - p_B \\ &= .051 - .045 \\ &= .006. \end{aligned}$$

Furthermore, since p equals .0481 (the proportion of defectives in both samples), we find

$$\begin{aligned} \hat{\sigma}_d^2 &= p(1 - p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \\ &= (.0481)(.9519)\left(\frac{1}{5,000} + \frac{1}{4,800}\right) \\ &= .00001868 \end{aligned}$$

and

$$\begin{aligned} \hat{\sigma}_d &= \sqrt{.00001868} \\ &= .0043 \end{aligned}$$

Thus, we find the critical value is

$$\begin{aligned}d^* &= 2.33\hat{\sigma}_d \\&= 2.33(.0043) \\&= .010.\end{aligned}$$

Since the value of d is .006 and this is less than the critical value, $d^* = .010$, our decision rule calls for the use of method A.

Our brief discussion of the statistical decision procedure for comparing two population proportions should indicate the general nature of such a procedure. However, you should recognize that we have treated herewith only the simplest situation—the formulation of a rule with n set directly and one risk set for D equal to zero. Technical complications, beyond the scope of this book, arise when objectives other than these are set. However, the basic ideas remain the same.

12.4 Decisions Involving the Comparison of Two Populations —Design

Let us further analyze the illustration discussed in the last section. The problem involved the collection of two samples—one sample of 5,000 inlays produced under one method of oiling the metal for a punch-press operation and another sample of 4,800 inlays produced under a second method of oiling the metal. In making these observations as a basis for decision, the industrial engineer must be careful to insure that both samples are random samples and that both samples are independently drawn. The theory for comparing two population proportions presented in the last section rests on these assumptions, and so does the theory for comparing two population means presented in Section 12.2.

Because it requires total randomization, the experimental design that we have considered to this point in our discussion is called a *completely randomized experiment*. An alternative experimental design which may have greater efficiency (that is, be less risky) than a completely randomized experiment is called a *matched samples* or a *paired samples* design.

Example 12.10 A garment manufacturer is considering the purchase of new sewing machines to expedite the manufacture of dresses. The decision on whether or not to buy the new machines is to be made by a comparison of the average times taken to sew the seams on a garment.

An experiment is to be designed to provide a basis for decision. One possible design would be a completely randomized experiment.

However, suppose that it were known that the choice of operators for the machines would influence the results. Then the experiment should be designed in such a way as to control this source of variation. If it were not, the decision procedure would be too risky. For example, suppose that experienced operators were inadvertently used only with the old machines, while apprentice operators were used with the new machines. Would a smaller average time on the old machines indicate that they were better than the new ones? Not necessarily. Variation in results due to differences in workers must be controlled, and this may be accomplished through the use of a matched samples design.

Thus, suppose that a matched sample design were adopted with workers as the basis for matching because they are a potential source of variation in the results. In such an experiment, each worker in the sample would be observed once on the new machine and once on the old machine. Thus, if 25 workers were used in an experiment, 25 pairs of observations would be made available. One possible set of results from such an experiment follows:

Worker	(Time Taken to Sew Seams—in Seconds)		
	Old Machine	New Machine	Difference
1.....	223	190	33
2.....	218	191	27
3.....	263	230	33
4.....	249	232	17
5.....	267	229	38
6.....	261	226	35
7.....	475	411	64
8.....	382	355	27
9.....	262	226	36
10.....	258	223	35
11.....	355	325	30
12.....	314	301	13
13.....	224	192	32
14.....	225	202	23
15.....	269	240	29
16.....	284	243	41
17.....	249	216	33
18.....	256	223	33
19.....	271	247	24
20.....	314	285	29
21.....	211	200	11
22.....	282	250	32
23.....	301	278	23
24.....	295	280	15
25.....	410	360	50
Average.....	284.72	254.20	30.52

You will notice that the difference between the two means is

$$284.72 - 254.20 = 30.52$$

and that this equals the average of the differences in the last column.

Before carrying out the matched samples experiment illustrated above, it would be necessary to formulate a statistical decision rule, and this, in turn, requires that objectives in terms of tolerable risks be specified. Thus, let us assume that the sample size for the afore-mentioned decision problem was set directly at 25 observations per machine. In addition, suppose that an analysis of the pertinent cost data indicates that if, on the average, the new machines are 35 seconds quicker than the old machines, only a .01 risk of not buying the new machines should be tolerated. Let μ_o be the average time, in seconds, necessary to sew the seams on the old machines and μ_N the average time on the new machines. Furthermore, let

$$D = \mu_o - \mu_N.$$

It has been specified that if $D = 35$, the risk of retaining the old machines must be only .01.

Since n has been set directly, only the critical value, d^* , need be determined for the statistical decision rule. This value should divide values of

$$d = \bar{x}_o - \bar{x}_N$$

into two regions, as follows:

RETAIN OLD MACHINES	BUY NEW MACHINES
d^*	

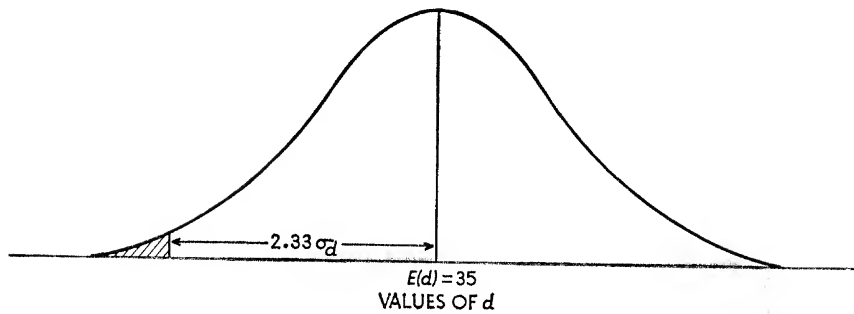
VALUES OF d

To find the value of d^* we must consider the sampling distribution of d for the type of experimental design and the type of sampling being utilized. This sampling distribution is, for sufficiently large samples, approximately normal. Its mean, $E(d)$ equals D . Its standard deviation—the standard error of d , namely σ_d —while ordinarily unknown a priori can be estimated from a sample by the following formula:

$$\hat{\sigma}_d^2 = \frac{\sum (d_i - d)^2}{n(n-1)} \quad (i = 1, 2, \dots, n),$$

where the values of d_i represent individual differences in the sample.

Let us apply this theory to find the critical value for the sewing machine decision problem. If $D = 35$, the sampling distribution of d would appear as follows:



You will notice that a tail area of .01—corresponding to the specified risk of retaining the old machines if $D = 35$ —has been shaded. Also notice that this implies d^* is 2.33 standard errors less than the value of $E(d)$ —that

$$\begin{aligned} d^* &= E(d) - 2.33\sigma_d \\ &= 35 - 2.33\sigma_d. \end{aligned}$$

This value may be computed after the sample is drawn and σ_d is estimated. For a value of d less than d^* , the decision should be to retain the old machines. For a value of d equal to or greater than d^* , the decision should be to purchase the new machines.

Example 12.11 Reconsider the data given in Example 12.10. The value of the pertinent statistic, d , is

$$\begin{aligned} d &= \bar{x}_O - \bar{x}_N \\ &= 30.52. \end{aligned}$$

To compare this to d^* we must first estimate σ_d . Hence, note that

$$\begin{aligned} \hat{\sigma}_d^2 &= \frac{\Sigma(d_i - d)^2}{n(n-1)} \quad (i = 1, 2, \dots, n) \\ &= \frac{1}{25 \times 24} [(33 - 30.52)^2 + (27 - 30.52)^2 + \dots + (50 - 30.52)^2] \\ &= \frac{3,062.24}{600} \\ &= 5.10. \end{aligned}$$

Therefore,

$$\hat{\sigma}_d = \sqrt{5.10} = 2.26$$

and

$$\begin{aligned} d^* &= 35 - 2.33\sigma_d \\ &= 35 - 2.33(2.26) \\ &= 29.73. \end{aligned}$$

Since d equals 30.25 and, hence, is greater than the critical value, d^* , the decision should be to buy the new machines.

The utility of the matched samples design which we have briefly discussed in this section is that it generally results in smaller risks, for the same amount of experimentation, than does a completely randomized experiment. (Alternately we might say that generally the matched samples design requires less experimentation to achieve the same reduction in risks.)

Any gain of the matched samples design arises from the grouping employed. In our sewing machine illustration, the observations were grouped by workers since it was logical to assume that times of operation differed from worker to worker. Grouping on some other bases might yield even better results. In any event, the principle to follow is to use groupings within which observations are likely to be less variable than within the population as a whole. (With this principle in mind it is worthwhile for a student to compare the matched samples design with stratified sampling—both rest on essentially the same basic idea.)

12.5 Decisions Involving the Comparison of Several Populations

Many statistical decision problems require experiments to compare not only two but several populations with respect to pertinent parameters. There are many different problems which have this general requirement, and each of these calls for its own particular type of statistical experiment. While a study of most of these is beyond the scope of this textbook, we shall consider the most common type of problem. This is one for which one must decide whether or not the populations under consideration have the same arithmetic means.

Example 12.12 A company's operations and analysis group has, among its duties, the job of rating employees. These ratings are used to check on the accuracy of ratings made by departmental foremen.

There are several efficiency analysts who rate employees in the group. In designing the procedure to check ratings (assigning analysts to departments, and so forth) the supervisor wishes to know whether or not there is any significant variation among his analysts in assigning ratings. If there is not, he need not worry about which analyst to assign to which department; if they do differ, he must take this into account in designing the rating procedure.

The last example illustrates a very common type of statistical problem. Let us define it statistically. Several populations are to be compared. Each population represents the totality of ratings that could be assigned by an analyst. (These populations are infinite

since the decision problem centers on the rating processes.) The problem is to determine whether or not the means of the populations are equal; for example, to decide whether or not the average ratings that would be assigned by analysts (in many, many trials) are equal. If they are, one course of action is called for; if they differ considerably, another course of action is appropriate. In short, we have the following:

Alternative	State of the World	
	$\mu_1 = \mu_2 = \mu_3 = \dots$	$\mu_1 \neq \mu_2 \neq \mu_3 \neq \dots$
Assume analysts differ	Type I error	Correct decision
Assume analysts similar	Correct decision	Type II error

The simplest design for a statistical study of this type of problem would require that a simple random sample of ratings be observed for each analyst. Thus, the supervisor might select a sample of 5 employees for analyst A to rate, another sample of 5 employees for analyst B to rate, and so forth. Let us assume there are 6 analysts. Then, 6 independently drawn samples of 5 employee ratings would be observed. In general, this statistical design would require k independently drawn simple random samples of size n .

Alternatives to this simple design will be considered in the next section. Here let us see how sample observations drawn in the manner sketched can be analyzed to provide a basis for decision. For reasons that will become obvious once one understands the methods employed, this method of statistical analysis is called the *analysis of variance*.

As an aid in discussing the analysis of variance, let us regard one possible outcome of the experiment designed above. Thus, consider the following data which represent 6 samples of 5 employee ratings. The ratings are on an efficiency scale which ranges from 0 to 50.

RATINGS OF EMPLOYEES BY SIX ANALYSTS

	A	B	C	D	E	F
	42	43	38	44	32	30
	38	42	28	44	42	43
	27	45	33	38	38	40
	49	38	40	41	29	25
	44	44	42	27	30	44
$\bar{x}_i$:	40.0	42.4	36.2	38.8	34.2	36.4

To analyze data such as these by the standard analysis of variance techniques, one must make the following assumptions:

1. The samples are simple random samples,
2. The samples are independent,
3. The populations are normal, and
4. The populations all have the same variance.

If any of these assumptions cannot be realistically made, the standard techniques that we shall consider must be modified. Recall, now, that the purpose of applying the analysis of variance is to decide whether or not the several population means are equal. Under the standard assumptions of normality and equal variances, this is equivalent to the decision on whether or not the several populations are identical. (If several normal populations all have the same variance, they can only differ if their means differ.)

The analysis of variance technique rests on a three-step procedure:

1. The variation *within* each of the several samples is calculated and averaged over all samples. This measure reflects variation in the populations.

2. The variation *between* sample means is calculated. This measure reflects variation in the populations and variation (if any exists) between the means of the populations.

3. The two measures of variation are compared to determine whether or not any variation between the means of the populations is indicated.

Let's consider each of these steps in detail. First, note that we can measure the variability within each sample by the sample variance, s^2 . For the i th sample this shall be denoted by s_i^2 . Thus, the population variance for each population may be estimated as

$$\hat{\sigma}_i^2 = \frac{n}{n-1} s_i^2.$$

However, since we assume in analysis of variance that the variability is the same for each population, we can obtain a better estimate of the population variance by averaging all of the values of $\hat{\sigma}_i^2$. In general, with k samples we average the k values of $\hat{\sigma}_i^2$. If we denote this estimated population variance determined on the basis of variability within the samples by $\hat{\sigma}_w^2$, we have

$$\hat{\sigma}_w^2 = \frac{\sum \hat{\sigma}_i^2}{k} \quad (i = 1, 2, \dots, k).$$

Example 12.13 Let us compute $\hat{\sigma}_w^2$ for the employee ratings given above. First, recall that the observations in the first sample are 42, 38, 27, 49, and 44 and that their mean is 40. Hence,

$$s_1^2 = \frac{(42 - 40)^2 + (38 - 40)^2 + \dots + (44 - 40)^2}{5}$$

$$= \frac{274}{5}$$

and

$$\hat{\sigma}_1^2 = \frac{n}{n-1} s_1^2$$

$$= \frac{5}{4} \cdot \frac{274}{5}$$

$$= 68.5.$$

Similarly, for the other samples we also can find estimates of σ_i^2 . Thus, we have:

<i>Analyst</i>	$\hat{\sigma}_i^2$
A.....	68.5
B.....	7.3
C.....	32.2
D.....	49.7
E.....	31.2
F.....	71.3

The value of $\hat{\sigma}_w^2$ is the average of these 6 values, or

$$\hat{\sigma}_w^2 = 43.37.$$

This is an estimate of the variation in the population of ratings for each analyst, based on the variability within each sample.

Now, consider the variation between sample means. A measure of this variation is the variance among all possible sample means. We call this measure $\sigma_{\bar{x}}^2$.

An estimator of $\sigma_{\bar{x}}^2$ is:

$$\hat{\sigma}_{\bar{x}}^2 = \frac{k}{k-1} \cdot \frac{\sum (\bar{x}_i - \bar{\bar{x}})^2}{k} \quad (i = 1, 2, \dots, k)$$

where $\bar{\bar{x}}$ is the over-all mean of the k samples.

Let's apply this estimator to our ratings illustration.

Example 12.14 The 6 sample means given above are:

40.0, 42.4, 36.2, 38.8, 34.2, 36.4.

The over-all mean is 38.0. Hence we shall estimate $\sigma_{\bar{x}}^2$ as

$$\hat{\sigma}_{\bar{x}}^2 = \frac{6}{5} \cdot \frac{(40.0 - 38)^2 + (42.4 - 38)^2 + \dots + (36.4 - 38)^2}{6}$$

$$= \frac{6}{5} \cdot \frac{44.24}{6}$$

$$= 8.848.$$

The measure computed in the last illustration reflects variation in the populations and variation (if any exists) between the means of the populations. Suppose, for the time being, that the hypothesis that the means are all equal is true. If so σ^2 should reflect only population variation. In fact, the 6 sample means could be regarded as coming from the same population. In this case we know

$$\sigma_x^2 = \frac{\sigma^2}{n}.$$

Since $n = 5$ for our problem, we have

$$\sigma_x^2 = \frac{\sigma^2}{5}$$

or

$$5 \times \sigma_x^2 = \sigma^2.$$

Hence, if all population means are equal, we may estimate σ^2 on the basis of the variation between sample means as follows:

$$\begin{aligned}\hat{\sigma}_b^2 &= 5\hat{\sigma}_x^2 \\ &= 5 \times 8.848 \\ &= 44.24,\end{aligned}$$

where $\hat{\sigma}_b^2$ is an estimate of σ^2 based on the variation between sample means.

The last step in our analysis is to compare $\hat{\sigma}_w^2$ and $\hat{\sigma}_b^2$ by taking the ratio of the latter to the former. Thus, for our illustration the variance ratio is

$$\begin{aligned}\frac{\hat{\sigma}_b^2}{\hat{\sigma}_w^2} &= \frac{44.24}{43.37} \\ &= 1.02.\end{aligned}$$

If all population means are equal, this ratio should, on the average, be approximately equal to one (as it is here). However, if they differ, $\hat{\sigma}_b^2$ should be larger than σ^2 on the average and, hence, larger than $\hat{\sigma}_w^2$ on the average. In this case we should expect to find the variance ratio significantly greater than one.

The variance ratio, as our discussion has indicated, is the pertinent statistic upon which the choice of a course of action should rest. High values indicate the means differ; values in the neighborhood of one or lower indicate they do not.

To formulate a statistical decision rule for the type of problem considered in this section requires some consideration of the sampling distribution of the variance ratio. This distribution is not nor-

mal. It follows the F distribution, a sampling distribution that is not considered in this textbook. Thus, here we cannot find for given levels of risk the critical value of the variance ratio and, hence, we cannot further discuss the formulation of a decision rule. The purpose of the preceding discussion has been to discuss the logic of the analysis of variance.

In conclusion, we may note that a study to compare several populations may involve sampling from each and every population or it may involve sampling from a sample of populations. For example, in the analysts-ratings illustration discussed in this section, we have assumed that a sample of ratings is observed for each and every analyst—that every population being compared is sampled. This is called a *fixed model* situation by statisticians. If, however, there were many analysts and a sample of ratings had been observed for only a sample of analysts, the problem would have presented a *random model* situation. This situation involves the comparison of several populations on the basis of only a sample of the populations.

12.6 Decisions Involving the Comparison of Several Populations—Design

In the last section we considered the simplest design of a statistical study to compare several populations. In short, this design required k simple random samples of size n to be observed independently. At times a more efficient design is possible—that is, at times it is possible to reduce the risks involved in a study by modifying its design.

Example 12.15 Reconsider the illustration used throughout the last section. Six samples of 5 employees were drawn and assigned to the efficiency analysts. The resulting information was used as a basis for deciding whether or not the analysts assigned ratings which, on the average, were equal.

An alternative, undoubtedly more efficient, experimental design would require that only *one* set of 5 employees be drawn. Each analyst then would be required to rate each of the 5 employees.

The control of employee variation would reduce the possible variation between sample means and, hence, would reduce the risks of observing significantly different average ratings in an experiment when actually the populations were similar. Ordinarily, the only disadvantage of this alternative design is that it requires a slightly more complicated analysis of variance.

As originally formulated, the design of the study of analysts' ratings required the analysis of variation arising only from one factor

—analysts. The problem was, in fact, to decide whether or not analysts did vary. As modified in Example 12.15, the problem required the consideration of variation arising from two factors—employees as well as analysts. In the first case we have a one-factor experiment, while in the second case we have a two-factor experiment. It also is possible to have three-factor or four-factor or m -factor experiments, though the more factors one considers, the more complex the analysis.

There are two reasons for controlling more than one factor in an experiment. First, as already noted, the experimental design may be more efficient in that the risks of an incorrect decision are reduced. Secondly, it is possible that the purpose of the study is to provide information not only about one factor but about two or more.

Example 12.16 The research department of a manufacturing company is developing a new adhesive tape. The uses for which the tape is designed dictate that its elasticity is an important quality characteristic. Preliminary tests on the tape as it has thus far been developed indicate that it is too variable. Thus, an experiment is to be conducted to decide which factors contribute to this variability and, hence, which factors must be further investigated, modified, or otherwise controlled.

Among the factors to be examined as sources of potential variation in the experiment are the following 5 factors. Thus, the experiment would be called a 5-factor experiment.

1. *Machines* used to produce the tape.
2. *Speeds* at which the machines are operating.
3. *Grades* of the raw material used.
4. *Suppliers* of the raw material.
5. *Temperatures* at which the tape's elasticity is tested.

Not only should such an experiment provide information about whether or not each of these factors influences the tape's elasticity but it also should provide information about how each combination of these factors influences it. For example, it may well be that while neither machine nor speed in themselves influence quality, certain speeds on certain machines do interact to cause differences in the quality of the tape. Thus, good statistical design must provide information about *interactions* as well as direct effects.

In connection with the last illustration, it should be noted that even today many experiments both in industry and out are based on the erroneous idea that in any experiment all but one factor of interest should be kept as constant as possible. This is a one-at-a-time experimental method. It usually is a naïve and inefficient procedure, however, because it requires several experiments to produce

even some of the information that one good statistically designed experiment could provide.

In concluding this discussion let us consider one other illustration to indicate different types of experimental designs and the importance of gearing the experimental design to the problem at hand.

Example 12.17³ A supermarket chain feels that the proper packaging of potatoes might increase sales. Therefore, it decides to experiment with 4 types of packaging. These represent the one factor of interest in the experiment. However, other factors that might be controlled in order to provide a more efficient experiment include the location of the store at which potatoes are sold and the day of the week.

Thus, the supermarket has 4 stores (A, B, C, and D) at different locations. Management also decides to limit the first packaging experiment to Monday through Thursday. Below we shall consider and compare 3 experimental designs which differ in whether or not location and day of sales are controlled.

As a basis for this comparison we shall deal with a simple situation and some simple numerical observations. Thus, suppose that packaging has no effect on sales; that is, that the means of the 4 populations of sales with different packaging are equal. However, we shall assume that sales for different days of the week and in different stores do vary and, on the average, they are as follows irrespective of packaging:

EXPECTED SALES OF POTATOES
(Hundreds of Pounds)

Day of the Week	Store			
	A	B	C	D
Monday	16	17	27	18
Tuesday	19	29	22	17
Wednesday	31	26	28	28
Thursday	23	15	27	17

First consider the use of a completely randomized design in the packaging problem. For this design we would decide on the type of packaging to use in a given store at a given time in a completely random manner. Suppose that each form of packaging is to be used 4 times. The form of packaging may be assigned in a random manner by assigning numbers 1-4 to method I; 5-8 to method II; 9-12 to method III; and 13-16 to method IV. Suppose then, that a table

³ Based on Max E. Brunk and Walter T. Federer, "Experimental Designs and Probability Sampling in Marketing Research," *Journal of the American Statistical Association*, Vol. XLVII (1953), pp. 440-52.

of random numbers is consulted and the methods of packaging are assigned to the stores for the days of the week as follows:

COMPLETELY RANDOMIZED DESIGN

Day of Week	Store			
	A	B	C	D
Monday.....	III	I	II	IV
Tuesday.....	II	II	I	III
Wednesday.....	II	I	IV	III
Thursday.....	I	III	IV	IV

If the observations indicated earlier are made, we would find that the sales for each method of packaging are as follows:

SALES
(In Hundreds of Pounds)

Method:	I	II	III	IV
	23	14	16	28
	17	31	15	27
	26	29	17	18
	<u>22</u>	<u>27</u>	<u>28</u>	<u>17</u>
Average:	22.0	26.5	19.0	22.5

The range in the means of 7.5 hundred pounds (from 19.0 to 26.5 hundred pounds) is a rough indicator of the difference in effect of the various methods of packaging. The randomizing process has ensured that despite the effect of different days of the week and store, the variability for a given method of packaging is due to chance. We could, therefore, apply the analysis of variance technique discussed in Section 12.5 to discover if the average effects of the methods of packaging differ.

Even though the effect of day of the week and store on sales is randomized, they still obviously contribute to the difference in sales observed for each method of packaging. On the basis of chance alone, packaging method II, for example, was never used on Thursday nor was packaging method IV tried in store A.

Next, consider a two-factor or randomized block design, as it is called. We can eliminate the effect of day of the week on the difference in average sales by ensuring that each method of packaging is used during every day of the week. In this case, the days of the weeks are called blocks and the method of packaging to be used in each store is determined in a random manner. Numbers 1-4 can be assigned to methods I-IV, respectively. Numbers 1-4 are drawn at

random for each day. The first number determined the packaging method for store A, the second for store B, etc. Suppose that this is done and the results are as follows:

RANDOMIZED BLOCK DESIGN

Day of Week	Store			
	A	B	C	D
Monday.....	II	I	IV	III
Tuesday.....	I	IV	II	III
Wednesday.....	IV	III	II	I
Thursday.....	I	IV	II	III

If the observations indicated earlier are made, we would find the sales for each method of packaging are as follows:

Method:	I	II	III	IV
	19	16	26	31
	23	22	18	29
	17	28	17	15
	<u>28</u>	<u>27</u>	<u>17</u>	<u>27</u>
Average:	21.75	24.25	19.5	25.5

Again the range of the means of 6 hundred pounds (from 19.5 to 25.5 hundred pounds) is a rough indicator of the effect of varying the method of packaging. The difference is only 6 hundred pounds since we have eliminated the effect of different days of the week.

We could have used the randomized block design to eliminate the effect of different stores. The stores would have been considered as blocks, and each method of packaging, selected in a random manner, used on one day during the Monday through Thursday period.

Lastly, consider the Latin Square design for which it is possible to eliminate the effect of different days of the week and different stores at the same time. With this design each method of packaging is tried in each store and during each day. One systematic Latin Square arrangement appears below. You will notice that one method of packaging always occurs along a diagonal:

LATIN SQUARE DESIGN

Day of Week	Store			
	A	B	C	D
Monday.....	I	IV	III	II
Tuesday.....	II	I	IV	III
Wednesday.....	III	II	I	IV
Thursday.....	IV	III	II	I

If we now make the pertinent observations given earlier, we find the following:

SALES				
(In Hundreds of Pounds)				
Method:	I	II	III	IV
	16	19	31	23
	29	26	15	17
	28	27	27	22
	<u>17</u>	<u>18</u>	<u>17</u>	<u>28</u>
Average:	22.5	22.5	22.5	22.5

You will now observe that the average sales for each method of packaging are identical. Thus, when the effect of the difference in stores and difference in day of the week has been removed, no difference in average sales for each method of packaging is revealed. In other words, there appears to be no relationship between the method of packaging potatoes and the sales of potatoes.

Hence, the importance of the proper experimental design is obvious. If the completely randomized or randomized block design had been used, we might have concluded incorrectly that the method of packaging has an impact upon sales of potatoes.

Of course, the above data have been arranged to get the indicated (and extreme) results—to show how the different designs could result in different conclusions. Hence, certain important aspects of experimental design follow from these illustrations. The completely randomized design is adequate only where other factors are known to exercise a very minor impact on the characteristic of interest. The randomized block design is adequate when only one other factor exerts a sizable influence on the characteristic of interest. A Latin Square design may be used when two other factors are known to have an impact.

12.7 You Should Now Know That

A comparative experiment requires experimentation as a basis for comparing properties of two or more populations.

The statistical design of an experiment is a plan for collecting observations.

In planning an experiment it is always necessary to relate decisions to possible outcomes.

A common type of decision problem calls for the comparison of two populations because the consequences of taking an action depend upon the difference in the means of the populations.

A statistical decision rule for such a problem should specify the sample size, the sample allocation, the method of selecting the sample, and the critical value.

Some comparative experiments are necessary in order to provide partial information on the difference between two population proportions.

Usually matched or paired samples will yield more efficient results than a completely randomized experimental design.

Many statistical decision problems may require experimentation to compare several populations.

A useful set of statistical techniques applicable in many of these situations is called the analysis of variance.

One experimental design for comparing several populations is the completely randomized design.

Other designs which involve controlling one other and two other factors, respectively, are the randomized block design and the Latin Square design.

12.8 You Should Now Be Able to Solve

1. Critically discuss the following statement: The proper design of an experiment requires that all factors, except the one being examined, be held constant.

2. In each of the following problem situations, suggest an experimental design. Explain carefully and discuss the reasons for your choice.

- a) One of two machines must be adopted for a certain manufacturing operation. The one which results in a lower proportion of defective product should be chosen. In addition to the possibility that machines turn out different product, it is thought that the grade of raw material used affects the quality of the product. Assume that there are 5 grades of raw material used in the operation.
- b) In designing a procedure for testing the melting point of a newly developed alloy, a researcher wishes to know whether or not any of the following three factors result, on the average, in different readings: thermometers used, analysts making the readings, and equipment employed in melting the metal. Assume that 5 thermometers, 3 analysts, and 2 sets of equipment are available for use.
- c) A taste-testing experiment is to be conducted to decide between two methods of preparing a frozen vegetable. Several judges will be used to rate the taste of the vegetable.

3. The maintenance department of a concern must choose between 2 batteries for use in the firm's operations. A sample of both

types is to be tested to failure to provide a basis for decision. A total sample of 50 batteries is to be used in the experiment. If the difference in average time-to-failure between battery A and battery B is 30 minutes, then a risk of using Type B, equal to only .01, is desirable. Based on the objectives set, formulate a statistical decision rule.

4. The experiment sketched in Question 3 is conducted. The following times-to-failure (in minutes) are observed. Apply the rule formulated above.

A			B		
122	158	115	127	86	92
119	157	185	154	82	92
116	148	138	105	161	98
178	161	153	101	83	125
103	136	140	127	142	163
193	118	102	149	83	178
178	188	195	89	90	119
123	109	135	160	155	166
	112			95	

5. An advertising agency plans an experiment to compare the effectiveness of an advertisement in a local newspaper with the effectiveness of a direct-mail campaign in small towns. In 5 towns, an ad is placed in the local paper. In another 5 towns, a direct-mail campaign is conducted. Two simple random samples of 2,500 residents each are selected. Residents are asked whether they remember the advertisement. Let π_A and π_D be the proportions of persons who remember the ad in each group of towns. It is decided that if $\pi_A = \pi_D$, then a risk of only .02 of using the direct-mail campaign on a nation-wide basis should be taken.

Formulate a statistical decision rule for this problem based on the objectives given above.

6. The experiment discussed in Question 5 is conducted. It is found that p_A is .2125, while $p_D = .2800$. Apply the rule formulated above.

7. The personnel department of a large insurance company wishes to test the effectiveness of a suggested program of special stenographic instruction. The program would be given to all secretarial personnel in the company. A preliminary experiment is to be designed to provide a basis for deciding whether or not to give the program. A sample of 25 employees (one class) is to be selected, and for each employee the number of words per minute taken in dicta-

tion is to be observed. After they complete the program, another stenographic test is to be given, and again speed is to be observed.

Let μ_B be the population average for employees before the program and μ_A the average after the program. Furthermore, let $D = \mu_A - \mu_B$. The personnel administrator of the company specifies that if the training results in no improvement on the average (i.e., if $D = 0$), only a .01 risk of adopting the program should be assumed. Based on these objectives, formulate a statistical decision rule.

8. The experiment indicated in Question 7 is conducted. The following observations (words per minute) are observed. Apply the rule formulated above.

Employee	Before Instruction	After Instruction
1.....	88	92
2.....	108	107
3.....	106	109
4.....	112	113
5.....	107	118
6.....	108	103
7.....	111	109
8.....	126	129
9.....	129	129
10.....	110	113
11.....	103	105
12.....	92	95
13.....	115	122
14.....	107	103
15.....	102	115
16.....	99	102
17.....	105	116
18.....	107	121
19.....	121	110
20.....	128	130
21.....	115	115
22.....	102	103
23.....	108	107
24.....	113	118
25.....	135	135

12.9 You Will Also Find That

For other brief introductory discussions of experimental design see:

- DUNCAN, ACHESON J. *Quality Control and Industrial Statistics*, pp. 735-64. Rev. ed. Homewood, Ill.: Richard D. Irwin, Inc., 1959.
- TIPPETT, L. H. C. *Technological Applications of Statistics*, pp. 159-84. New York: John Wiley & Sons, Inc., 1950.

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics: A New Approach*, pp. 479-83. Glencoe, Ill.: The Free Press, 1956.

More extended discussions will be found in the following key references:

COCHRAN, WILLIAM G., and COX, GERTRUDE M. *Experimental Designs*. New York: John Wiley & Sons, Inc., 1950.

FISHER, R. A. *The Design of Experiments*. 5th ed. Edinburgh, London: Oliver & Boyd, 1949.

Decisions by Association

13.1 Introduction

In this chapter we are going to cover some statistical theory and some statistical applications which are essentially different from those in the preceding chapters. As the title of this chapter tells you, we are going to study decisions by association. What does this mean? In what ways is this topic different from others we have studied? These questions can be answered in an elementary way with reference to a familiar situation.

Example 13.1 The main character in this example is a typical American housewife, Mrs. Smith. The scene of the example is the neighborhood store. Mrs. Smith is considering whether or not to buy some grapes. She has a decision problem: to buy or not to buy.

The grocer is a good neighbor. Mrs. Smith is a good customer. The grocer, in fact, always allows Mrs. Smith to sample a few (but not too many) of the grapes as a basis for her decision. In our terminology, Mrs. Smith is allowed to make her decision on the basis of observations from the population of grapes on which she is making a decision.

In a broad way, the above example describes the type of statistical decision problem that we have been discussing in many different business connections. Almost all of our problems, including the grape-buying decision, have dealt with a specific set of elementary units. In addition, the pertinent characteristics of these elements have made up a relatively accessible population. Certainly, in all our examples, samples could be drawn from these populations and observations made on the characteristics of interest.

Let's go back to Mrs. Smith at the grocery.

Example 13.2 Mrs. Smith, having decided on the grapes (she bought a pound), eyes the cantaloupes. Unfortunately, she cannot

taste the cantaloupe to test its flavor. (Even as good a customer as Mrs. Smith is not permitted such liberties.) She, therefore, smells the cantaloupe, balances it, and squeezes it. On the basis of observations of smell, balance, and hardness, she reaches her decision as to the taste of the melon.

The decision-making experiment in this case is quite different from the procedure followed in deciding on the purchase of grapes. To reach a decision on the lot of grapes, grapes are sampled because the elementary units are accessible and the pertinent characteristic (taste) of the grapes sampled can be observed easily. In the case of the melon, however, a slice of the melon is not accessible and the characteristic of the melon cannot be observed. Mrs. Smith, therefore, has to make a decision by association. On the basis of past experience she is confident that if a melon smells sweet, is balanced perfectly, and is neither too hard nor too soft, its taste is pleasant. Thus, Mrs. Smith relies on the results of her observations of characteristics that she has found are associated with tasty melons.

To make decisions by association, we try to predict one variable (taste) on the basis of other variables (smell, balance, hardness). In general, the variable which we try to predict is called the *dependent variable*. The variables which are the basis for the prediction are called *independent variables*. In this chapter we shall limit our discussion primarily to situations which involve the relationship between a dependent variable and *one* independent variable. This is called a *simple* relationship. In such cases, a statistical study deals with a *bivariate* population; that is, a population which includes a pair of observations for each elementary unit. If we were to consider a relationship between a dependent variable and two or more independent variables, we should have to consider a *multiple* relationship and, hence, a *multivariate* population.

Decisions by association are not uncommon in the world of business. Let us consider a few illustrations.

Example 13.3 A personnel department has to decide on the acceptability of candidates for a position. It would be too time consuming and expensive to give each applicant an on-the-job trial. On the basis of prior experience, the personnel department has observed that job performance is related to a candidate's score on a series of tests. It, therefore, attempts to predict performance on the basis of associated variables—test scores.

Example 13.4 A firm manufacturing airplanes must weld joints on the wing of the plane. The wing must be able to withstand shear-

ing forces while in flight. The test to determine shearing strength of a joint, however, is destructive in nature. Therefore, an associated variable—the weld diameter—is observed to predict shearing strength.

Example 13.5 An appliance manufacturer desires to predict his company's sales for the coming year. It has been found that disposable personal income is associated with the sales of the firm. The firm, therefore, predicts its sales on the basis of estimates of disposable personal income made by economists.

Example 13.6 A cost accountant wishes to predict the cost of material handling for the coming year. He uses an associated variable—direct labor cost—to make his decision on the estimate.

In each of the above illustrations, because of considerations of cost, time, or availability of data, one variable is estimated on the basis of its relationship to an associated variable.

13.2 Two Aspects of Association

There are two aspects to problems involving decisions by association. One aspect concerns itself with the *degree of relationship*. This refers to the extent to which decisions are enhanced by taking into consideration an associated variable. The second aspect concerns itself with the *nature of the relationship*. This refers to the functional relationship between the variables which enables us to predict the value of a dependent variable from the value of an independent variable.

Example 13.7 In the field of education the degree of association between arithmetic and reading is of interest in order to reach decisions on methods of teaching. If it is found that there is a high degree of association between arithmetic and reading ability, a pupil's reading ability should be considered in diagnosing retardation in arithmetic. The interest here is not to predict a pupil's arithmetic ability on the basis of his reading ability but to reach a decision on whether arithmetic ability might be enhanced by reading ability. Therefore, the degree of relationship is relevant.

In contrast, all the business illustrations in Examples 13.3 through 13.6 are illustrations of problems where the nature of the functional relationship is paramount. Thus, to reach a decision on sales for the following year, on the basis of disposable income, it is necessary to know the nature of the relationship between the variables.

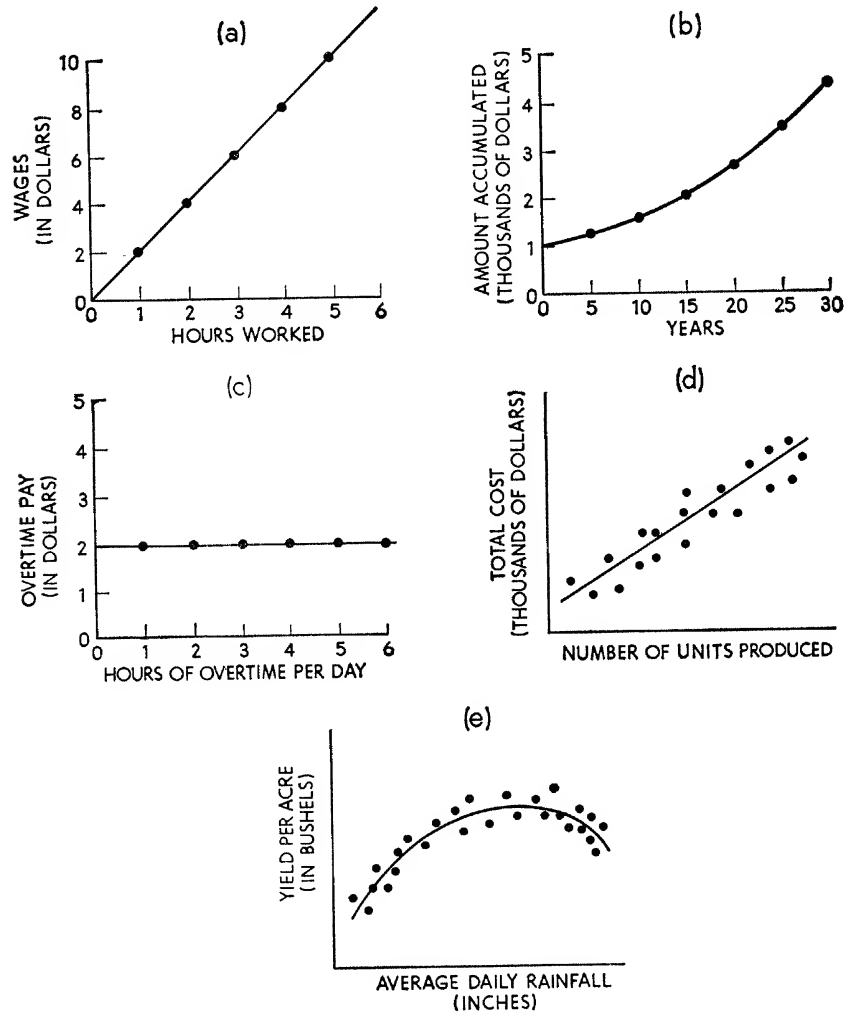
Statisticians refer to problems concerning the degree of relationship as *correlation* problems and to those involving the nature of re-

relationship as *regression* problems. We can gain further insight into the nature of correlation and regression problems by considering the diagrams in Figure 13.1.

Figure 13.1(a) illustrates the relationship between wages earned

Figure 13.1

SOME RELATIONSHIPS BETWEEN PAIRS OF ASSOCIATED VARIABLES



and hours worked when the wage rate is \$2.00 per hour. The nature of the relationship is described by a straight line whose equation is $Y = 2X$, where Y represents wages earned and X hours worked. The degree of relationship or correlation in this case is perfect since once

the number of hours worked is known, the amount a worker has earned is determined exactly.

Figure 13.1(b) represents the relationship between the amount to which an investment of \$1,000 (compounded annually at 5 per cent) will accumulate and the number of years that the money is invested. Here, again, the degree of relationship is perfect. However, in this case, the nature of relationship is not described by a straight line but by a curve. The curve is an exponential curve and is expressed by the equation: $Y = 1,000 (1.05)^x$

Figure 13.1(c) illustrates the nature of the relationship between overtime pay and the number of hours overtime worked a day for an agency that allows only a flat \$2.00 for overtime. Obviously, there is no relationship between hours of overtime worked and pay since no matter how many hours of overtime an employee works, the pay remains the same. Graphically, the absence of relationship is represented by a straight line parallel to the X-axis.

Figure 13.1(d) represents the relationship between total cost and number of items produced by a firm. The degree of relationship in this case lies somewhere between the two extremes of perfect relationship and no relationship. Decisions can be enhanced by the use of the associated variable in such instances, but the value of one variable cannot be precisely determined by the value of the associated variable. A straight line seems to offer a description of the nature of the general relationship. This illustration represents a situation that is typical of association between variables encountered in business problems.

Figure 13.1(e) also represents a relationship that is not perfect, but in this case the nature of the relationship is better represented by a curve rather than a straight line. This relationship tells us that yield per acre increases with the amount of rainfall and then starts to decrease. This is logical for if rainfall were to increase indefinitely, crops would be flooded and there would be no yield.

As we have seen, regression relationships can be represented by straight lines or by curves. Relationships described by straight lines are referred to as *linear*; those represented by curves as *curvilinear*. However, in this text we shall limit our discussion to linear relationships.

The diagrams that were drawn in the previous illustration are called *scatter diagrams*. Scatter diagrams are very useful devices for investigating the nature of the relationship between variables. By convention, the independent variable is considered the X-variable,

and it is plotted on the horizontal scale; that is, on the X -axis. The dependent variable is considered the Y -variable, and it is plotted on the vertical scale or Y -axis.

In any problem involving a decision by association, a scatter diagram will provide a clue to the nature of and the degree of the relationship. Many erroneous and needless computations can be avoided by this simple device. Thus, the use of a straight line relationship when a curve would offer the proper description would lead to unreliable results, and perhaps to an incorrect decision.

13.3 Association and Causation

A word of caution is needed in the interpretation of the meaning of association. The mere fact that two variables are associated does not imply that one variable is a *cause* of the second variable. Thus, the fact that the smell and the taste of most melons are related does not mean a sweet smell causes a sweet taste. Neither is the fact that an individual who scores high on a test also performs well on the job mean that a high test score is a cause of good performance.

The presence or absence of association can be demonstrated through the use of statistics. However, the determination of whether a cause-and-effect relationship exists depends on knowledge of the substantive field to which the data apply. Thus, it can be shown that the incidence of heart disease is associated with the level of consumer prices in the United States. However, this does not prove that an increasing price level is a cause of a higher incidence of heart disease. To prove cause and effect, a researcher would have to connect, on the basis of subject matter, high prices and the physiology of the heart. Although the statistical association is real, its interpretation as evidence of a cause-and-effect relationship is erroneous.

Why is it that associations can be shown even though cause-and-effect relationships are not present? One reason for statistical association between variables that are not related logically is that they are both related to other factors. Thus, the relationship between consumer prices and the incidence of heart disease occurs because both of these variables have been increasing over a period of time. The passage of time is the third factor implicitly responsible for the statistical relationship.

Example 13.8 During the past 25 years there has been a close relationship between teachers' salaries and liquor consumption in the United States. This relationship can be explained by the fact that both variables are related to the level of income in the country. In-

creases (or decreases) in teachers' salaries are associated with increases (or decreases) in national income. Similarly, increases (or decreases) in liquor consumption are associated with increases (or decreases) in national income. The relationship between teachers' salaries and liquor consumption, therefore, is due to their relationship to a third factor—national income.

Can you now explain: Why high grades are associated with low hours of study? Why the number of storks' nests in an English village was related to the size of its population?

A second reason for the observation of a relationship between variables even though no cause-and-effect relationship exists is a small-size sample. Thus, it is possible to draw a sample which exhibits some association even though there is no relationship between the variables in the population. In a later section of this chapter we will describe a test to see whether the association we observe could be merely a chance relationship.

A third reason for an observed relationship may be due to the fact that both variables contain a common element. This is especially true if ratios with common elements are associated with one another. If two ratios have the same denominator, then a small denominator tends to make both ratios large and a large denominator tends to make both ratios small. Conversely, if two ratios have the same numerator, a small numerator will tend to make both ratios large. It is not surprising, therefore, to find that some correlation exists between the net profits/net assets ratios and the net profits/net sales ratios for firms within an industry.

The expressions, "spurious correlation" and "nonsense correlation," often are used to describe situations where a high degree of association exists without any cause-and-effect relationship. However, it is not the degree of association that is spurious—it is the explanation of the high degree of relationship. Therefore, great care must be taken in explaining the meaning of a close statistical association between variables.

If a cause-and-effect relationship is to exist, however, there must be association between the variables. In recent years, there has been much debate as to whether cigarette smoking is a cause of lung cancer. To demonstrate the possibility of a cause-and-effect relationship it was essential to demonstrate that the two variables were associated. Had this association not been demonstrated, it would not have been possible to support a claim that smoking is a cause of lung cancer. However, the existence of the statistical association does not

prove that a cause-and-effect relationship exists. The interpretation of the observed association is the basis of the present conflict.

We can sum up the discussion in this section with the statement: *While causation implies association, association does not imply causation.*

13.4 A Population Model for Regression Problems

We now direct our attention to one of the aspects of decisions by association: regression, the nature of the relationship. As a basis for this discussion, let us reconsider some subject matter from an earlier chapter. In order to estimate the relative frequency of observations within given limits in a population we ordinarily must assume some population model (such as the normal curve). For example, suppose that we are studying the specifications for the diameters of ball bearings produced by a given process. Further suppose that we estimate that the diameters average .15 inches and that their standard deviation is .002 inches. If we then assume that the normal curve is an adequate population model, we may state: "About 95 per cent of the ball bearings produced will have diameters between .146 and .154 inches." The basis for this statistical prediction is the assumption of a normal population model.

In a regression problem we are interested in predicting one variable on the basis of its relationship to a second variable. In order to make statistical predictions we again must make certain assumptions about the population. These assumptions constitute a *bivariate population model*.

The particular model with which we shall be concerned in this section is called the *simple linear-regression model*. There are many other models which can be used to describe the association between two variables. The simple linear-regression model, however, is one which frequently applies to the association between variables and is one for which the techniques for making predictions have been developed rather fully.

We can introduce the characteristics of this model through an illustration.

Example 13.9 A manufacturer of detergents is interested in predicting the behavior of a particular compound under varying temperatures. This will help him to decide on whether the new compound is good enough to be produced. The relationship between detergency and temperature is an important factor. To wash delicate items of clothing, water at low temperatures is used. The detergent must, therefore, have sufficient detergency at low temperatures to have a

cleansing effect. Coarser items of apparel are washed at higher temperatures. The detergency at these high temperatures should not be so excessive as to impede the proper operation of the washing machine, or to interfere with the washing and rinsing functions.

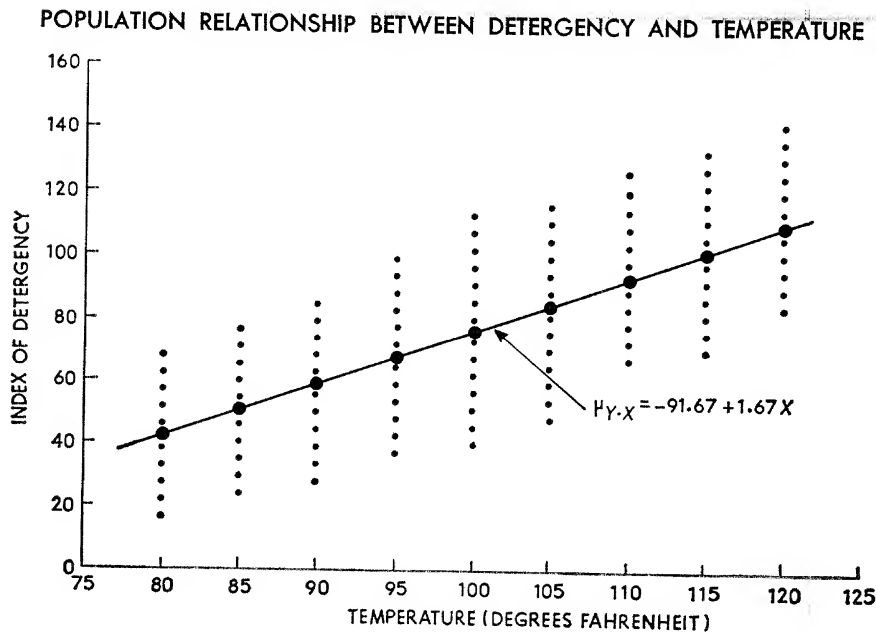
The manufacturer desires to determine on the basis of a sample study the nature of the relationship between the amount of detergency and temperature so that he can predict the behavior of his product under varying temperatures.

The population in this problem is bivariate. For each elementary unit (a fixed amount of detergent) we must observe two characteristics—its detergency and the temperature to which it is subjected. Let us consider the assumptions which this bivariate population must satisfy for the simple linear-regression model to apply.

The first assumption is that the relationship in the population is linear. That is, a straight line describes the relationship between detergency and temperature. In Figure 13.2 the straight line describes the relationship between the two variables. We call this line the *population regression line of Y on X*. Detergency is the Y-variable since it is the variable we desire to predict. Temperature is the X-variable since it is the associated variable to be used as a basis for prediction. Hence, the line represents the regression of detergency on temperature.

A second assumption is that for a given value of X, the mean of

Figure 13.2



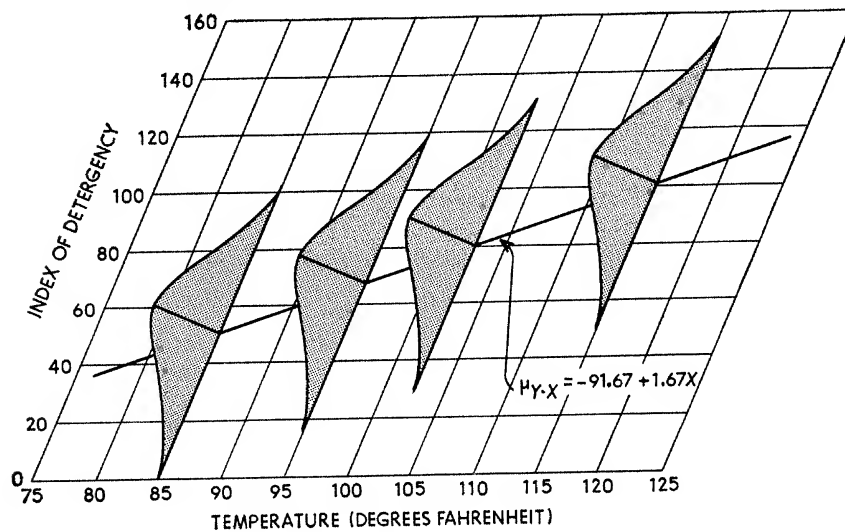
the Y -values falls on the line of regression. If we refer to the chart, we note that for a temperature of 85 degrees the average index of detergency is 50. Similarly, the average index of detergency at a temperature of 115 degrees is 100.

A third assumption is that the variance (and standard deviation) of the Y -values about the line is the same for each X -value. This means that for any given temperature the scatter of the index of detergency about the line should be the same.

A fourth assumption is that the Y -values for a given X -value are normally distributed. For example, consider the distribution of indexes of detergency at a temperature of 85 degrees. To apply the model under consideration we must assume that the indexes of detergency at this temperature are normally distributed. The same assumption also must be made at all other temperatures considered. Graphically, we may depict the assumption for our illustration by Figure 13.3.

Figure 13.3

POPULATION RELATIONSHIP BETWEEN DETERGENCY AND TEMPERATURE



The assumptions of a simple linear-regression model are summarized below:

1. The population relationship is linear.
2. The mean of the Y -values for a given X -value lies on the regression line.

3. The variances of the Y -values about the regression line are constant for all X -values.
4. The distribution of Y -values for any X -value is normal.

Let us now develop some of the properties of the simple linear-regression model more fully and formulate them in symbolic terms. First of all, we know that in order for the model to be applicable, the Y -values associated with a given X -value must be normally distributed. Thus each set of Y -values comprises a normal curve with its own mean and standard deviation. We may designate the means by the symbol

$$\mu_{Y \cdot X},$$

which is read "the mean of the Y -values for a given X ." We already have observed that for a temperature of 85 degrees, the average Y -value (index of detergency) is 50. This indicates that

$$\mu_{Y \cdot 85} = 50.$$

Similarly, the fact that at a temperature of 115 degrees the average Y -value is 100 may be indicated by

$$\mu_{Y \cdot 115} = 100.$$

In order to meet the conditions of the simple linear-regression model, all of the values of $\mu_{Y \cdot X}$ must lie on the regression line. This line, which relates $\mu_{Y \cdot X}$ and X , may be expressed, in general, as follows:

$$\mu_{Y \cdot X} = A + BX.$$

For the detergency-temperature illustration we shall assume that the equation is

$$\mu_{Y \cdot X} = -91.67 + 1.67X.$$

You may recognize the general form of this relationship as the algebraic equation for a straight line. A and B are constants which determine a particular line; often A is called the Y -intercept and B is called the slope of the line. Thus, A is the value of $\mu_{Y \cdot X}$ when X equals zero. However, A need not have any practical significance. In our problem it is not necessary to consider a value of zero for X (temperature). The value of A (-91.67), therefore, has no meaningful interpretation. It has only mathematical significance—together with B it locates the regression line.

The slope, B , refers to the change in $\mu_{Y \cdot X}$ per unit of increase in X . In our reference example, B represents the change in the aver-

age index of detergency as temperature increases 1 degree. Since $B = 1.67$, this means that $\mu_{Y \cdot X}$ increases 1.67 units per degree increase in temperature.

When there is no relationship between the independent and dependent variable, the value of B is zero. In other words, there is no change in $\mu_{Y \cdot X}$ per unit of increase in X .

The two constants in the regression equation have been symbolized by the capital letters, A and B , because they are population parameters. (The Greek letters, α and β , have not been employed because they already have been used in this textbook in another connection.) In the next section we shall consider the problem of estimating A and B on the basis of sample observations. For the present, however, note that once we have determined A and B , we can find, from the regression line, the average value of the detergency index for any given temperature.

Example 13.10 We have assumed that for the bivariate population of detergency and temperature

$$\mu_{Y \cdot X} = -91.67 + 1.67X.$$

Hence to determine the average index of detergency if the detergent were exposed many times to temperatures of 90 degrees, we let $X = 90$ and find

$$\begin{aligned}\mu_{Y \cdot 90} &= -91.67 + 1.67X \\ &= -91.67 + 1.67(90) \\ &= -91.67 + 150.3 \\ &= 58.6\end{aligned}$$

The regression line that we have considered in this section describes the relationship for the bivariate population where temperatures range from 80 to 120 degrees. We have assumed nothing about the relationship that exists outside of this range. Hence, we should not use the regression line to determine the average index of detergency for temperatures, for example, of 40 or 140 degrees. Of course, we could substitute X -values of 40 and 140 degrees in the regression equation and determine the corresponding average index of detergency. However, this use of the regression line would be on the basis of our judgment that the same relationship would exist at these temperatures.

Let us now consider the standard deviation of the Y -values for a given X -value. That is, let us consider the scatter about their mean ($\mu_{Y \cdot X}$) of possible indexes of detergency for a given X -value. Since the distribution of Y -values for a given X -value is normal, if

we know their standard deviation the extent of scatter can be determined. The formula for the standard deviation of the Y -values for a given X -value is

$$\sigma_{Y.X} = \sqrt{\frac{\sum(Y - \mu_{Y.X})^2}{N}}$$

where Y is an observed Y -value for a given X . The standard deviation of the Y -values for a given X is called the *standard deviation of regression* or the standard error of estimate. The latter term may be confusing because $\sigma_{Y.X}$ is a property of a population, and the term "standard error" usually is associated with errors in sampling procedures. Hence, we shall use the term "standard deviation of regression."

Example 13.11 Let us assume that the standard deviation of regression for the index of detergency is 6 units. Then we know that approximately 68 per cent of the time, if a detergent is used with water at 115 degrees, the index of detergency will fall between

$$\begin{array}{ccc} \mu_{Y.115} - \sigma_{Y.115} & \text{and} & \mu_{Y.115} + \sigma_{Y.115} , \\ 100 - 6 & \text{and} & 100 + 6 , \\ 94 & \text{and} & 106 . \end{array}$$

Similarly, about 95 times in 100, the index will fall between

$$\begin{array}{ccc} \mu_{Y.115} - 2\sigma_{Y.115} & \text{and} & \mu_{Y.115} + 2\sigma_{Y.115} , \\ 100 - 12 & \text{and} & 100 + 12 , \\ 88 & \text{and} & 112 . \end{array}$$

Outside of what limits will the index fall about 27 times in 10,000?

13.5 The Estimated Regression Line

In almost every practical problem requiring decisions by association, the population line of regression and the population standard deviation of regression are unknown. They must be estimated on the basis of the results of a statistical study.

The standard sampling procedure in a regression problem for which the simple linear-regression model is applicable is to observe random samples of Y -values for fixed values of the X -variable. Thus, in the detergency problem the detergent might be tested several times at each of several fixed temperatures. For example, four tests might be performed at a temperature of 85 degrees, four other tests made at a temperature of 100 degrees, and four other tests made at a temperature of 120 degrees. The outcomes (observed indexes of detergency) would be assumed to constitute three random samples of four observations.

You will observe that in this experimental procedure we assume that the X -variable is completely controlled and that the X -values do not vary due to chance. They are fixed, or as is sometimes said, the values of X are error-free. For each fixed value of X , however, a random sample of Y -values is observed.

How do we know that the Y -values for a given X -value constitute a random sample? How do we know that these Y -values are independent of each other—that the magnitude of one does not influence the magnitude of a second? The answer to these questions is similar to the answer to a similar question considered in our study of statistical quality control. In that field of application it was essential that a process behave like a random process in order for the control chart to be used. However, this could be assumed in most applications because experience had indicated that most production processes behaved as random processes. Similarly, here it seems reasonable to assume that outcomes of the detergency test should not influence one another—that they should constitute random samples from the population.

Suppose that samples of Y -values are observed for various fixed values of the X -variable. A and B would then be estimated on the basis of the sample data. The estimate of A is denoted as a , the estimate of B as b . Hence, the equation of the line as estimated from the sample is

$$\hat{\mu}_{Y.X} = a + bX.$$

On other occasions, we have indicated that statisticians very often use unbiased estimates of parameters. We select, for a and b , values which are unbiased estimates of A and B , respectively. The formula for b is:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2},$$

where x and y represent values of the variables in the sample and n is the sample size. The formula for a is:

$$a = \bar{y} - b\bar{x}.$$

Since we have washed our previous illustration clean, let us turn to a new example.

Example 13.12 A large mail-order house feels that there is an association between the weight of the mail it receives and the volume of orders to be filled. It desires to investigate the matter and to determine the nature of the relationship, if an association does exist. A knowledge of the number of orders early in the day will be

of tremendous help to the organization in planning its daily schedule of operations. The need for such planning is apparent when one realizes that a mail-order house handles from 1,000 to 4,000 orders in a 10- to 15- or 20-minute cycle. This means that anywhere from 1,000 to 4,000 orders are being processed, assembled, packed, and shipped through the same operational channels during each cycle. Any overlapping of these cycles can result in complete chaos.

To illustrate the calculations necessary to estimate the regression line, let us assume that a sample of 25 observations is selected within a range of X -values from 200 to 600 pounds of mail. Five observations are made for 200 pounds, five for 300 pounds, and so forth. The results are as follows:

Observation	Weight of Mail (Hundreds of Pounds)	Number of Orders (in Thousands)
1.....	2.0	6.4
2.....	2.0	8.0
3.....	2.0	7.2
4.....	2.0	7.5
5.....	2.0	6.9
6.....	3.0	10.9
7.....	3.0	10.3
8.....	3.0	9.5
9.....	3.0	9.7
10.....	3.0	10.6
11.....	4.0	12.5
12.....	4.0	12.9
13.....	4.0	14.0
14.....	4.0	13.8
15.....	4.0	12.8
16.....	5.0	16.5
17.....	5.0	17.1
18.....	5.0	15.4
19.....	5.0	16.2
20.....	5.0	15.8
21.....	6.0	19.0
22.....	6.0	19.4
23.....	6.0	19.1
24.....	6.0	18.5
25.....	6.0	20.0

We first construct a scatter diagram to investigate the nature of the relationship. This is illustrated in Figure 13.4. You will observe that each of the plotted points in the chart has been identified by a number to facilitate references to the original data. The scatter diagram indicates that a straight line may be used to describe the nature of the relationship between the two variables. In Example 13.13 the procedure for estimating the regression line is shown.

Example 13.13 The worksheet needed to estimate the regression line is arranged as follows:

WORK SHEET FOR COMPUTATION OF REGRESSION LINE

Observation	Weight of Mail (Hundreds of Pounds) x	Number of Orders (in Thousands) y	xy	x^2
1.....	2.0	6.4	12.8	4
2.....	2.0	8.0	16.0	4
3.....	2.0	7.2	14.4	4
4.....	2.0	7.5	15.0	4
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
22.....	6.0	19.4	116.4	36
23.....	6.0	19.1	114.6	36
24.....	6.0	18.5	111.0	36
25.....	6.0	20.0	120.0	36
Total.....	100.0	330.0	1,470.0	450

$$\begin{aligned}
 n &= 25 & \Sigma xy &= 1,470.0 \\
 \Sigma x &= 100.0 & \Sigma x^2 &= 450 \\
 \Sigma y &= 330.0
 \end{aligned}$$

$$\begin{aligned}
 b &= \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \\
 &= \frac{25(1,470.0) - (100.0)(330.0)}{25(450) - (100.0)^2} \\
 &= \frac{36,750 - 33,000}{11,250 - 10,000} \\
 &= \frac{3,750}{1,250} \\
 &= 3.00 \text{ thousand orders per 100 pounds}
 \end{aligned}$$

This value of b means that for each increase of 100 pounds in weight we estimate that the number of orders will increase by 3,000.

$$\begin{aligned}
 a &= \bar{y} - b\bar{x} \\
 &= \frac{\Sigma y}{n} - \frac{b\Sigma x}{n} \\
 &= \frac{330}{25} - \frac{3(100)}{25} \\
 &= 13.2 - 12 \\
 &= 1.20 \text{ thousand orders}
 \end{aligned}$$

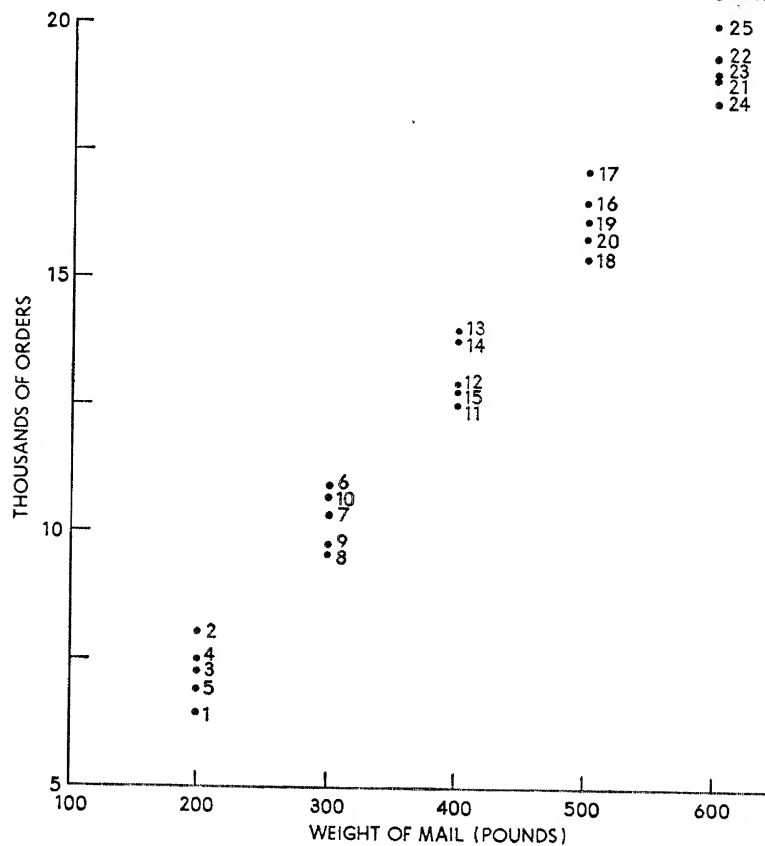
Hence the equation of the regression line is

$$\hat{\mu}_{Y.X} = 1.20 + 3.00X.$$

Since X ranges from 200 to 600 pounds in our sample, this regression equation should be used only when the daily weight of mail falls within this range.

Figure 13.4

RELATIONSHIP BETWEEN NUMBER OF ORDERS AND WEIGHT OF MAIL



The Least-Squares Method. The regression line that we have computed also can be derived by a technique statisticians call the *least-squares method*. For the simple linear-regression model, the unbiased estimates of A and B are identical with the values we would derive by the least-squares method. Historically, the least-squares method preceded the development of the concept of unbiased estimates. The technique was introduced by the French mathematician, Adrien Legendre, more than 150 years ago.

The least-squares method fits a regression line to the data that satisfies the following criterion: The sum of the squared deviations about the line is a minimum.

To determine the values of a and b for the least-squares equation, a set of normal equations must be solved. These equations are:

$$(1) \Sigma y = na + b\Sigma x$$

$$(2) \Sigma xy = a\Sigma x + b\Sigma x^2$$

If we were to solve these equations for a and b for the data in Example 13.13, we would get the same answers as before.

The Standard Deviation of Regression. We have stated previously that the standard deviation of regression in the population measures the variability of the Y -values about the population regression line. In a practical problem we must estimate the standard deviation of regression from the sample. It can be shown mathematically that an unbiased estimator for the variance of regression is

$$\hat{\sigma}_{Y \cdot X}^2 = \frac{\sum (y - \hat{\mu}_{Y \cdot X})^2}{n - 2}.$$

An estimator for the standard deviation of regression is, therefore,

$$\hat{\sigma}_{Y \cdot X} = \sqrt{\frac{\sum (y - \hat{\mu}_{Y \cdot X})^2}{n - 2}}.$$

Example 13.14 We estimate the value of $\sigma_{Y \cdot X}$ below. Note that the estimated number of orders is determined by substituting the actual weight in the regression equation:

$$\hat{\mu}_{Y \cdot X} = 1.20 + 3.00X.$$

WORK SHEET FOR COMPUTATION OF STANDARD ERROR OF ESTIMATE

Observation	Weight of Mail (Hundreds of Pounds) x	Observed Number of Orders (in Thousands) y	Estimated Number of Orders (in Thousands) $\hat{\mu}_{Y \cdot X}$	$y - \hat{\mu}_{Y \cdot X}$	$(y - \hat{\mu}_{Y \cdot X})^2$
1....	2.0	6.4	7.2	-0.8	.64
2....	2.0	8.0	7.2	+0.8	.64
3....	2.0	7.2	7.2	0.0	.00
4....	2.0	7.5	7.2	+0.3	.09
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
22....	6.0	19.4	19.2	+0.2	.04
23....	6.0	19.1	19.2	-0.1	.01
24....	6.0	18.5	19.2	-0.7	.49
25....	6.0	20.2	19.2	+0.8	.64
Total..	100.0	330.0	330.0	0.0	7.52

$$\begin{aligned}\hat{\sigma}_{Y \cdot X} &= \sqrt{\frac{\sum (y - \mu_{Y \cdot X})^2}{n - 2}} \\ &= \sqrt{\frac{7.52}{23}} \\ &= .572 \text{ thousand orders}\end{aligned}$$

You will observe that the sum of the deviations about the line is equal to zero. This always should be so when unbiased estimators are used and offers a method of checking the accuracy of calcula-

tions. Also note that the sum of the estimated Y -values is equal to the sum of the actual values and, hence, the mean of the estimated values is equal to the mean of the actual Y -values. This provides another check on the calculations.

13.6 Sampling Error of Estimated Regression Line

We need a regression line to predict the value of a dependent variable on the basis of the value of an independent variable. For example, in the previous illustration the regression line is needed to predict the number of orders on the basis of the weight of the mail received. This prediction would enable a mail-order house to plan its activities for the day.

Predictions based on a regression line derived from a sample will be subject to two sources of error. In the first place, the relationship between the two variables may not be perfect. Such errors in predictions would occur even if the population regression line were known. They do not arise because of sampling. The size of this error is indicated by the standard deviation of regression. In the second place, the regression line is computed from a sample and, hence, does not necessarily show the true position of the population regression line. This section is devoted to a discussion of the nature of the sampling errors and their measurement.

The Sampling Error of b . The slope (b) of an estimated regression line is a random variable. Let us, therefore, consider the pattern of variability of b . If we were to draw a very large number of samples of n pairs of observations from a population that meets the assumptions of a simple linear-regression model, and if the values of X were the same for each of the samples, then the b values would be normally distributed and their mean would be equal to B . That is, the expected value of b is B . Furthermore, the pattern of variability of these b values could be determined if we knew the standard error of b . The formula for the standard error of b is

$$\sigma_b = \frac{\sigma_{Y \cdot X}}{\sqrt{\sum (x - \bar{x})^2}}.$$

You will recognize the numerator as the population standard deviation of regression. The denominator is the square root of a sum—the sum of the squared deviations of the X -values about their mean.

The formula reveals three facts about the standard error of the slope: (1) The smaller the standard deviation of regression, or the more perfect the relationship in the population, the smaller is the standard error of the slope. (2) The larger the sample size, the smaller is the standard error of the slope. This follows since the

larger the sample, the more $(x - \bar{x})^2$ terms are added. This increases the denominator and decreases σ_b . (3) The greater the variability of the x -values, the smaller is the standard error of b .

Since we do not know the population standard deviation of regression in any given problem, we must estimate its value from a sample. Hence, an estimator for σ_b may be written as

$$\hat{\sigma}_b = \frac{\hat{\sigma}_{Y \cdot X}}{\sqrt{\Sigma(x - \bar{x})^2}}.$$

The fact that the estimated value of the slope is not zero does not mean that a relationship necessarily exists between the variables in the population. A nonzero value of b might arise because of sampling error. To know with certainty whether a relationship does or does not exist, it would be necessary to know the value of the slope in the population. If B were equal to zero, no relationship would exist and the regression line would be useless to predict Y on the basis of X . If B were not equal to zero, a relationship would exist and the regression relationship might be useful. Since population data are not available, it is necessary to reach a decision on whether or not a relationship exists on the basis of sample data. The decision problem may be summarized as follows:

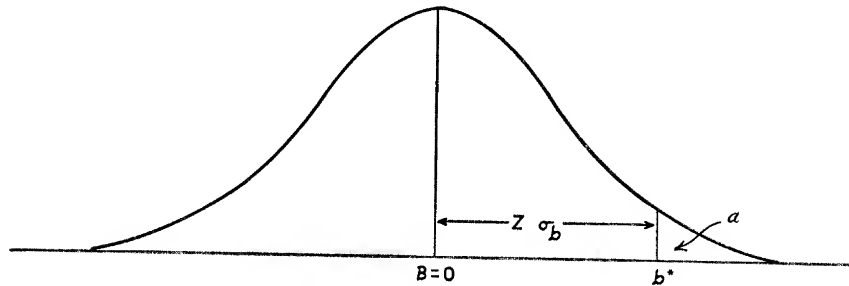
Action	State of the World	
	$B = 0$	$B \neq 0$
Consider line as basis for prediction	Incorrect decision	Correct decision
Discard line as basis for prediction	Correct decision	Incorrect decision

We shall consider situations in which the sample size is given. In such cases it is possible to control only one type of error—the error that we are most eager to avoid. This error usually is to use the regression line if no relationship exists. Thus, if we set a value for the α -risk, we can derive a decision rule for which the probability of using the regression line when no relationship exists is equal to α . As usual, we determine the appropriate value of α on the basis of the costs entailed in making an incorrect decision. In the mail-order house problem, the consequences of predicting the number of orders on the weight of mail, if no relationship between the variables exist, must be considered.

What we seek, therefore, is a decision rule to give us the desired protection against the use of the regression line to estimate Y on the basis of X if no relationship exists. The decision rule would have the

following form: If in a sample of n items from a bivariate population b is greater than b^* (the critical value), consider the regression line as a basis for prediction; if b is equal to or less than b^* , discard the regression line as a basis for prediction.

We have noted that the sampling distribution of b is normal. It is, therefore, necessary to determine a value of b^* so that the probability of obtaining values of b equal to or greater than b^* is α if B is equal to zero. This situation is illustrated below:



The diagram indicates that

$$b^* = 0 + z\sigma_b$$

or

$$b^* = z\sigma_b.$$

Example 13.15 Let us now derive a decision rule to determine whether or not to consider the use of the regression equation in Example 13.13 to estimate the number of orders on the basis of the weight of mail.

We set an α -risk of .05. We know that

$$\begin{aligned} b^* &= z\sigma_b \\ &= 1.64\sigma_b. \end{aligned}$$

(The normal deviate for an area of .05 in one tail is 1.64.) Now note that

$$\begin{aligned} \sigma_b &= \frac{\hat{\sigma}_{Y \cdot X}}{\sqrt{\sum (x - \bar{x})^2}} \\ &= \frac{\hat{\sigma}_{Y \cdot X}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}} \\ &= \frac{.572}{\sqrt{450 - \frac{(100)^2}{25}}} \\ &= \frac{.572}{\sqrt{50}} \\ &= .081. \end{aligned}$$

Hence,

$$\begin{aligned} b^* &= (1.64)(.081) \\ &= .133. \end{aligned}$$

The decision rule, therefore, is: If the value of the slope (b), estimated from a sample of 25 from a bivariate population of number of orders and weight of mail, is equal to or less than .133, discard the regression line; if b is greater than .133, do not discard the regression line.

Since b is equal to 3.00 (in Example 13.13), we should not discard the regression line as a basis for estimating number of orders on the basis of weight of mail.

Even though the decision is that the line should be considered as a basis for estimating the dependent variable upon the basis of the independent variable, we still must determine whether the estimates based on the regression line will be sufficiently precise to meet the requirements of a given problem situation. In terms of the mail-order house problem, we must determine whether the precision of the estimate of the number of orders, based on the weight of mail, is precise enough to make possible effective planning of a day's operations.

The Standard Error of an Estimated Mean. Since the estimated regression line is determined from a sample of observations, the estimates of the Y -intercept (A) and the slope (B) are subject to the possibility of sampling error. Therefore, the estimate of $\mu_{Y \cdot X}$ which depends upon the estimates of A and B will also be subject to the possibility of sampling error.

For simplicity in notation let $\bar{y}$ be an estimate of $\mu_{Y \cdot X}$ for a given X , where $\bar{y}$ is computed from an estimated regression equation. Thus,

$$\hat{\mu}_{Y \cdot X} = \bar{y} = a + bX.$$

We now shall consider the sampling error of this estimate for a given X -value. An estimator for the standard error is given by the following formula:

$$\sigma_{\bar{y}} = \sigma_{Y \cdot X} \sqrt{\frac{1}{n} + \frac{(X - \bar{x})^2}{\sum (x - \bar{x})^2}},$$

where $\sigma_{\bar{y}}$ is the standard error of the estimate of the average Y -value for a given X -value, denoted by X .

The first part of the expression under the square root sign reflects the error in a , and the second part indicates the error in b . You will observe that $\sigma_{\bar{y}}$ varies with X . It is at a minimum when

X is equal to the mean of the observed X -values. As the deviation of the value of X from $\bar{x}$ increases in either direction, $\sigma_{\bar{y}}$ increases.

Example 13.16 Let us now determine the standard error for the line of regression in the mail-order house problem. We have worked out in detail the calculation of the error in the average value of the number of orders ($\bar{y}$) when the weight of the mail is 500 pounds ($X = 5$). The errors for other values of X are given in the table that follows the computation.

$$\hat{\sigma}_{\bar{y}} = \hat{\sigma}_{Y \cdot X} \sqrt{\frac{1}{n} + \frac{(X - \bar{x})^2}{\Sigma(x - \bar{x})^2}}$$

The values of $\hat{\sigma}_{Y \cdot X}$ and $\Sigma(x - \bar{x})^2$ were computed previously as .572 (see Example 13.13) and 50 (see Example 13.15), respectively. Therefore,

$$\begin{aligned} \hat{\sigma}_{\bar{y}} &= .572 \sqrt{\frac{1}{25} + \frac{(5 - 4)^2}{50}} \\ &= .14 \text{ for } X = 5. \end{aligned}$$

The above result of .14 means that if many, many regression lines were computed for the same X -values, then the resulting average value of the number of orders for 500 pounds of mail would have a standard deviation of .14 thousand orders (140 orders).

Values of $\hat{\sigma}_{\bar{y}}$ for
Selected X -Values

X	$\hat{\sigma}_{\bar{y}}$
2.....	.20
3.....	.14
4.....	.11
5.....	.14
6.....	.20

Observe that the value of $\hat{\sigma}_{\bar{y}}$ is at a minimum when X is 4 (the mean of the X -values) and increases as we deviate from the mean in either direction.

The error in the line, as we have stated, represents a sampling error. As with all sampling errors, the error can be decreased by increasing the sample size. A glance at the formula for $\hat{\sigma}_{\bar{y}}$ will convince you of this. Both denominators, n and $\Sigma(x - \bar{x})^2$, increase as sample size increases and hence $\hat{\sigma}_{\bar{y}}$ decreases.

The Standard Error of an Estimated Y -Value. We have considered how the standard error for the *mean* of the Y -values for a given X -value can be computed. This error increases as X deviates from the average X -value. It also can be decreased by increasing the sample size. This standard error is the result of the use of a sam-

pling procedure. It approaches zero if larger and larger samples are taken.

We now turn to the standard error of an estimate of an *individual* Y -value for a given X -value. The standard error of this type of estimate is the result of two factors: (1) sampling, and (2) an imperfect relationship in the population. Thus, an estimate of an individual Y -value, even if the population were known, would be subject to some error—this we know is indicated by the population standard deviation of regression.

Let y be the symbol for an estimate of an individual Y -value. Such an estimate may be determined from the regression equation. That is, a point on the regression line will be used as an estimate. Hence, let

$$y = a + bX.$$

Note that the estimate of an individual Y -value is determined from the regression equation in exactly the same way that estimates of the *average* value of Y for a given X -value are determined. However, in the first case the standard error will be larger, for we are trying to predict an individual rather than an average value.

Thus, while the standard error of an estimate of an average Y -value may be estimated as

$$\hat{\sigma}_y = \hat{\sigma}_{Y \cdot X} \sqrt{\frac{1}{n} + \frac{(X - \bar{x})^2}{\sum (x - \bar{x})^2}},$$

an estimate of the standard error of an individual Y -value is

$$\hat{\sigma}_y = \hat{\sigma}_{Y \cdot X} \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{x})^2}{\sum (x - \bar{x})^2}}.$$

As you can see by inspection of these formulas, the latter always will be larger since it includes an extra one under the square root. Let's illustrate the difference by a numerical example.

Example 13.17 Let us continue with the mail-order problem. Suppose that 5 hundred pounds of mail ($X = 5$) are received on a given day. On the average this will mean

$$\begin{aligned}\hat{\mu}_{Y \cdot X} &= 1.20 + 3.00(5) \\ &= 16.2 \text{ thousand orders.}\end{aligned}$$

We also may estimate that on the particular day Y is 16.2 thousand orders; that is, $y = 16.2$.

These estimates are the same, but their standard errors are not. In Example 13.16 we found that

$$\hat{\sigma}_y = .14 \text{ thousand orders}$$

for the estimate of the average Y -value. The standard error of the individual estimate is, however,

$$\begin{aligned}\hat{\sigma}_y &= \hat{\sigma}_{Y.X} \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{x})^2}{\sum (x - \bar{x})^2}} \\ &= .572 \sqrt{1 + \frac{1}{25} + \frac{(5 - 4)^2}{50}} \\ &= .59 \text{ thousand orders.}\end{aligned}$$

Estimated values of σ_y for other values of X are included in the table below. You should compare them to the corresponding estimates of $\sigma_{\bar{y}}$ given in Example 13.16.

X	y	$\hat{\sigma}_y$
2.....	7.2	.61
3.....	10.2	.59
4.....	13.2	.58
5.....	16.2	.59
6.....	19.2	.61

Again you will notice that the standard error of the estimated Y -values is at a minimum when $X = 4$ hundred pounds, which is also the mean of the x -values in the sample. The standard error increases as X deviates from 4 in either direction.

Interpretation of Standard Errors. For sufficiently large samples a regression estimate of $\mu_{Y.X}$ or of an individual Y -value will be normally distributed. In such cases the standard errors of these estimates permit measurement of risks in the usual way—by reference to the normal curve. Thus, the risk that an estimate of $\mu_{Y.X}$ differs from the true value by $1.96\sigma_{\bar{y}}$ or more would be .05. The risk that it differs by $2.58\sigma_{\bar{y}}$ or more would be .01. Similarly, an estimate of an individual Y -value for a given X -value may be evaluated by reference to the normal curve if n is large. The risk of an error equal to or greater than $1.96\sigma_y$ in such an estimate would be .05, etc.

Example 13.18 In Example 13.17 we estimated $\mu_{Y.X}$ as 16.2 thousand orders for $X = 5$ hundred orders. We also computed the standard error of this estimate as

$$\hat{\sigma}_{\bar{y}} = .14 \text{ thousand orders.}$$

We now note that

$$1.96\hat{\sigma}_{\bar{y}} = .27 \text{ thousand orders.}$$

There is a risk of .05 that an estimate of $\mu_{Y.X}$ will be in error by .27 or more thousand orders (that is, by 270 or more orders). Similarly,

$$2.58\hat{\sigma}_{\bar{y}} = .36 \text{ thousand orders.}$$

Hence, there is a risk of .01 that an estimate of $\mu_{Y \cdot X}$ will be in error by 360 or more orders.

For estimates of $\mu_{Y \cdot X}$ at other levels of X , we find risks of exceeding the following errors:

X	Risk: .05 $1.96\hat{\sigma}_{\bar{y}}$	Risk: .01 $2.58\hat{\sigma}_{\bar{y}}$
2.....	.39	.52
3.....	.27	.36
4.....	.22	.28
5.....	.27	.36
6.....	.39	.52

These values are all interpreted in the way indicated above.

Example 13.19 In Example 13.17 we considered the value of 16.2 thousand orders as an estimate of an individual Y -value for $X = 5$. In this case, we estimated the standard error as

$$\hat{\sigma}_y = .59 \text{ thousand orders.}$$

Now note that

$$1.96\hat{\sigma}_y = 1.16 \text{ thousand orders.}$$

Thus, there is a risk of .05 that an estimate of Y will be in error by 1.16 or more thousand orders. Furthermore, since

$$2.58\hat{\sigma}_y = 1.52 \text{ thousand orders,}$$

there is a .01 risk that an estimate of Y will be in error by 1.52 or more thousand orders.

The errors in the estimate of Y -values that would be exceeded 5 times in 100 and 1 time in a 100 are shown below for some other values of X :

X	Risk: .05 $1.96\hat{\sigma}_y$	Risk: .01 $2.58\hat{\sigma}_y$
2.....	1.20	1.57
3.....	1.16	1.52
4.....	1.14	1.516
5.....	1.16	1.52
6.....	1.20	1.57

Compare these possible errors with those given in the last illustration. For a given level of risk and a given X -value, these are larger. Why?

The proper interpretation of the standard errors of estimates of average and individual Y -values is important because it enables us to evaluate the worth of a regression line as a basis for decisions by association.

Example 13.20 Let us assume that the mail-order house can tolerate an error of up to 2 thousand orders in any day's estimate of orders received. Further, let us suppose that the risk of an error equal to or greater than 2 thousand orders should be only .05. Does the regression equation meet the needs of the mail-order house?

The answer is yes, if $1.96\hat{\sigma}_y$ is less than 2 thousand orders for all possible values of X . We already have computed the values of $1.96\hat{\sigma}_y$ in Example 13.19. All are less than 2 thousand orders. The estimated regression equation is satisfactory.

Confidence Intervals. At times it may be preferable to use a regression equation to derive an interval estimate of $\mu_{Y.X}$ or of an individual Y -value, rather than compute point estimates. A confidence interval, as we have noted in an earlier chapter, is a range of values which we hope includes the quantity being estimated. It is constructed by a procedure which if repeatedly used would have a specified probability of including the quantity being estimated.

Consider the computation of an interval estimate for $\mu_{Y.X}$. A .95 confidence interval for $X = 5$ hundred pounds would be:

$$\begin{array}{rcl} \hat{\mu}_{Y.5} - 1.96\hat{\sigma}_{\hat{y}} & \text{to} & \hat{\mu}_{Y.5} + 1.96\hat{\sigma}_{\hat{y}}, \\ 16.2 - 1.96(.14) & \text{to} & 16.2 + 1.96(.14), \\ 16.2 - .27 & \text{to} & 16.2 + .27, \\ 15.93 & \text{to} & 16.47. \end{array}$$

Similarly, a .99 confidence interval for $X = 5$ hundred pounds would be:

$$\begin{array}{rcl} \hat{\mu}_{Y.5} - 2.58\hat{\sigma}_{\hat{y}} & \text{to} & \hat{\mu}_{Y.5} + 2.58\hat{\sigma}_{\hat{y}}, \\ 16.2 - 2.58(.14) & \text{to} & 16.2 + 2.58(.14), \\ 16.2 - .36 & \text{to} & 16.2 + .36, \\ 15.84 & \text{to} & 16.56. \end{array}$$

If many samples of 25 were drawn and intervals such as this computed, 99 per cent of the intervals would include the true value of $\mu_{Y.X}$ for $X = 5$.

The computation of an interval estimate for an individual Y -value is similar to the above calculations. However, we must use y rather than $\mu_{Y.X}$ and σ_y rather than $\sigma_{\hat{y}}$ in the formulas. Thus, a .95 confidence interval for $X = 5$ hundred pounds would be:

$$\begin{array}{rcl} y - 1.96\hat{\sigma}_y & \text{to} & y + 1.96\hat{\sigma}_y, \\ 16.2 - 1.96(.59) & \text{to} & 16.2 + 1.96(.59), \\ 16.2 - 1.16 & \text{to} & 16.2 + 1.16, \\ 15.04 & \text{to} & 17.36. \end{array}$$

The calculation of .95 confidence intervals for $\mu_{Y.X}$ and for an individual Y -value at other X -values is given on page 428.

X	$\hat{\mu}_{Y \cdot X}$	$1.96 \sigma_{\hat{y}}$	$\hat{\mu}_{Y \cdot X} - 1.96 \sigma_{\hat{y}}$ to $\hat{\mu}_{Y \cdot X} + 1.96 \sigma_{\hat{y}}$
2.....	7.2	.39	6.8 to 7.6
3.....	10.2	.27	9.9 to 10.5
4.....	13.2	.22	13.0 to 13.4
5.....	16.2	.27	16.0 to 16.5
6.....	19.2	.39	18.8 to 19.6

X	y	$1.96 \sigma_y$	$y - 1.96 \sigma_y$ to $y + 1.96 \sigma_y$
2.....	7.2	1.20	6.0 to 8.4
3.....	10.2	1.16	9.0 to 11.4
4.....	13.2	1.14	12.1 to 14.3
5.....	16.2	1.16	15.0 to 17.4
6.....	19.2	1.20	18.0 to 20.4

Figure 13.5 presents these results. Note that the confidence band for the estimates of Y is wider than that for the estimates of $\mu_{Y \cdot X}$. Also observe that both confidence bands are narrowest when $X = 4$, the mean of the X -values, and both get wider as the X -values depart from this mean.

13.7 Judgment Applications of Regression

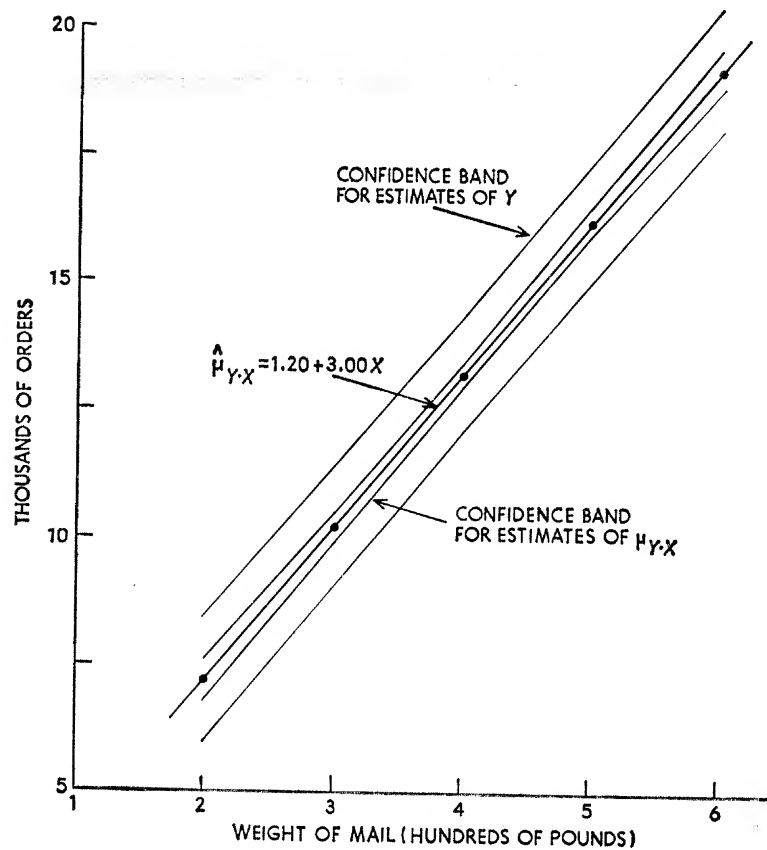
There are many occasions when a bivariate population of interest does not satisfy the conditions of the simple linear-regression model. For example, one assumption that is frequently violated is that the scatter of the Y -values about the regression line is normal. Another is that the standard deviation of regression is constant for all pertinent values of X . In these situations the estimation procedures described above are still valid, but the risks of errors in the estimates usually cannot be precisely determined.

Another problem arises in situations for which the X -values are not fixed for the study. Still another difficulty presents itself when the Y -values cannot be assumed to constitute simple random samples. In such cases none of the statistical theory and methods given above applies.

This does not mean that a regression line cannot be computed. In fact, the same formulas can be used. Any prediction based on the regression line in these instances, however, cannot be made on the basis of the statistical theory discussed above. Predictions based on such regression lines must rely on the judgment of the expert using this technique. A common application of regression analysis in this connection is to forecast sales of a company. Very often a variable such as gross national product or disposable personal income is used as the independent variable. However, the relationship of the varia-

Figure 13.5

.95 CONFIDENCE INTERVALS FOR $\mu_{Y \cdot X}$ AND FOR INDIVIDUAL Y-VALUES AT VARIOUS VALUES OF X

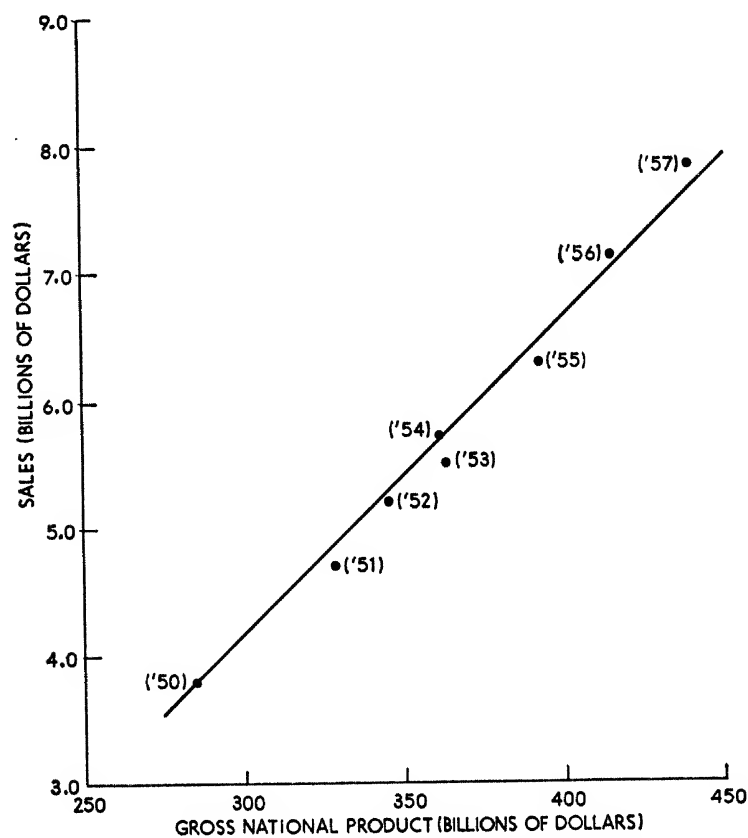


bles in the population does not satisfy the assumptions of the simple linear-regression model. For example, the X-values in such applications cannot be fixed for purposes of the study. In addition, the deviations of the annual sales from the line of regression are related to each other, casting suspicion on the assumption of randomness. As a rule, if sales of one year are high, they will be followed by relatively high sales during the next year.

Example 13.21 This illustration develops the relationship between the annual sales of the Standard Oil Company of New Jersey and the nation's output of goods and services as measured by gross national product. The scatter diagram for these two variables is shown in Figure 13.6. It indicates that the relationship between the variables may be described by a straight line. The regression line computed on page 431 also has been drawn on that chart.

Figure 13.6

RELATIONSHIP BETWEEN SALES OF STANDARD OIL COMPANY
OF NEW JERSEY AND GROSS NATIONAL PRODUCT



WORK SHEET TO DETERMINE REGRESSION LINE
SALES OF THE STANDARD OIL COMPANY OF NEW JERSEY
ON GROSS NATIONAL PRODUCT

Year	Sales (Billions of Dollars) y	GNP (Billions of Dollars) x	xy	x^2	y^2
1950	3.8	285	1,083.0	81,225	14.44
1951	4.7	328	1,541.6	107,584	22.09
1952	5.2	345	1,794.0	119,025	27.04
1953	5.5	363	1,996.5	131,769	30.25
1954	5.7	361	2,057.7	130,321	32.49
1955	6.3	392	2,469.6	153,664	39.69
1956	7.1	415	2,946.5	172,225	50.41
1957	7.8	440	3,432.0	193,600	60.84
Total	46.1	2,927	17,320.9	1,089,413	277.25

Sources: 1957 Annual Report, Standard Oil Company of New Jersey, and Survey of Current Business, July, 1958.

$$\begin{aligned}
 b &= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \\
 &= \frac{8(17,320.9) - 134,934.7}{8(1,089,413) - 8,567,329} \\
 &= \frac{138,567.2 - 134,934.7}{8,715,304 - 8,567,329} \\
 &= \frac{3,632.5}{148,075} \\
 &= .02453
 \end{aligned}$$

$$\begin{aligned}
 a &= \bar{y} - b\bar{x} \\
 &= 5.763 - (.02453)(365.88) \\
 &= 5.763 - 8.975 \\
 &= -3.21
 \end{aligned}$$

Hence, the equation of the regression line is

$$\hat{y} = -3.21 + .02453x.$$

This equation may be used to estimate sales ($\hat{y}$) for different levels of gross national product (x). For example, suppose that GNP is \$285 billion, as it was in 1950, we find

$$\begin{aligned}
 \hat{y} &= -3.21 + (.02453)(285) \\
 &= -3.21 + 6.99 \\
 &= \$3.8 \text{ billion.}
 \end{aligned}$$

To check on the "fit" of the regression line for the period 1950 to 1957, we may compute values of $\hat{y}$ for each year and measure the errors in the estimates. A judgment decision must then be made on the possible usefulness of the line for future estimates. The calculation of values of $\hat{y}$, as well as the computation of errors and percentage errors, is given below:

Year	Actual Sales y	Estimated Sales $\hat{y}$	Error $y - \hat{y}$	Percentage Error $\frac{y - \hat{y}}{\hat{y}} \times 100$
1950.....	3.8	3.8	0.0	0.0
1951.....	4.7	4.8	-0.1	-2.1
1952.....	5.2	5.3	-0.1	-1.9
1953.....	5.5	5.7	-0.2	-3.5
1954.....	5.7	5.6	+0.1	+1.8
1955.....	6.3	6.4	-0.1	-1.6
1956.....	7.1	6.9	+0.2	+2.9
1957.....	7.8	7.6	+0.2	+2.6
Total	46.1	46.1	0.0	

Observe that the sum of the estimated y -values is equal to the sum of the actual y -values and that, hence, the sum of the deviations of the actual values from the estimated values is zero. These relationships act as a check of the calculations.

The computations indicate that in the past, the use of the regression line to estimate sales would have resulted in errors that range from an underestimate of only 2.9 per cent to an overestimate of only 3.5 per cent. On the average, the percentage error is 2.1 per cent. If the company feels that it can tolerate errors of the magnitude observed, and that the same relationship between the variables will exist in the future, the line may be adopted as a means of forecasting. Thus, a forecast made in 1959 for 1960 would have to anticipate the value of gross national product for 1960, perhaps on the basis of an economist's judgment.

In the last example, there would be no point in trying to determine if a relationship exists by seeing if b differed from $B = 0$ on the basis of statistical theory. The theory does not apply here. Similarly, the standard error of an estimate of sales for a given level of GNP need not be computed since we could not attribute any statistical meaning to it. The risks of such estimates can be evaluated only on the basis of judgment.

The decision to use the line for purposes of prediction can depend, in part, on its past performance in predicting sales. If this past performance seems satisfactory and there are logical reasons to assume that this relationship will continue in the future, it might be decided to use the regression line as a basis for prediction. However, it is not possible to measure the risk of an error. Judgment, based on experience in the substantive field, and not on statistical theory, would be the basis of evaluating the prediction.

Work is being carried on by statisticians to develop more complicated models than we have used so that problems like the one discussed above might fall within the realm of statistical theory. Such developments, however, are beyond the scope of this text.

13.8 The Population Coefficient of Correlation

A regression equation, as we have stated, describes the nature of the relationship among variables. There are times, however, when we are not interested in the *nature of the relationship* but in the *degree of relationship* between the variables. One measure of the degree of relationship is called the *coefficient of correlation* (to be denoted by ρ , the Greek letter *rho*).

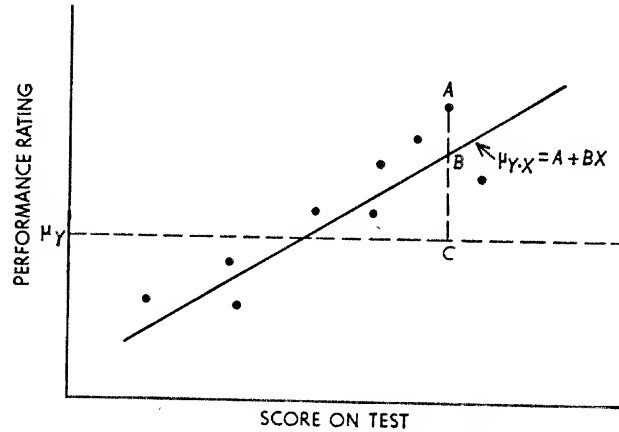
Example 13.22 The personnel director of a corporation wants to decide upon a set of psychological tests to administer to prospective candidates for an executive training program. He is to use the results of these tests in conjunction with other criteria (e.g., education, experience, personality, diction) to select candidates. He obviously

would prefer tests which are most closely associated with successful performance in the training course. He is, thus, interested in the degree of association between grades on a given test and performance.

The degree of such association may be measured by the coefficient of correlation. Tests which have the highest coefficients of correlation with performance are to be selected. Let us now assume that the population regression line is known. The line and some of the points about the line are plotted in Figure 13.7.

Figure 13.7

RELATION BETWEEN PERFORMANCE RATING AND TEST SCORE



If we desired to predict performance of candidates on the basis of past performance without reference to any associated variables, we might predict that a candidate's performance rating will be equal to the average performance rating in the population (μ_Y). The deviation (error) of an individual performance rating from the average could be expressed as $(Y - \mu_Y)$ and is represented by line AC in the above chart. This deviation (error) can be written as:

$$\overline{AC} = \overline{AB} + \overline{BC}$$

or

$$(Y - \mu_Y) = (Y - \mu_{Y.X}) + (\mu_{Y.X} - \mu_Y).$$

If we square the both sides of the equation, do the same for all points in the bivariate population, and then sum all of these squares, we obtain

$$\Sigma(Y - \mu_Y)^2 = \Sigma(Y - \mu_{Y.X})^2 + \Sigma(\mu_{Y.X} - \mu_Y)^2.$$

Let us examine these three expressions.

The term $\Sigma(Y - \mu_Y)^2$ is a measure of the variability of the Y -values about their mean and is called the *total sum of the squares*.

The term $\Sigma(\mu_{Y.X} - \mu_Y)^2$ is the sum of the squared deviations of the values of $\mu_{Y.X}$ about their mean. This term is a measure of the explained variability since it measures the variability of the $\mu_{Y.X}$ values which are determined on the basis of the regression line. It is referred to as the *explained sum of the squares*.

Finally, the term $\Sigma(Y - \mu_{Y.X})^2$ is a measure of the unexplained variability since it measures the difference between the observed Y -values and the value of Y we obtain on the basis of the regression relationship. It is called, the *unexplained sum of the squares*.

We, therefore, can state that

$$\begin{aligned} \text{TOTAL SUM OF THE SQUARES} = \\ \text{EXPLAINED SUM OF THE SQUARES} + \text{UNEXPLAINED SUM OF THE SQUARES.} \end{aligned}$$

We define ρ^2 (the square of the coefficient of correlation) as follows:

$$\rho^2 = \frac{\text{EXPLAINED SUM OF SQUARES}}{\text{TOTAL SUM OF SQUARES}}$$

or

$$\rho^2 = 1 - \frac{\text{UNEXPLAINED SUM OF SQUARES}}{\text{TOTAL SUM OF SQUARES}}.$$

Symbolically, we may write

$$\rho^2 = \frac{\Sigma(\mu_{Y.X} - \mu_Y)^2}{\Sigma(Y - \mu_Y)^2}$$

or

$$\rho^2 = 1 - \frac{\Sigma(Y - \mu_{Y.X})^2}{\Sigma(Y - \mu_Y)^2}.$$

In terms of our last illustration, ρ^2 represents the proportion of the variability in the population of performance ratings that is explained by variability in test scores.

If there were no relationship at all between the variables, Y would have the same average value for every X . In other words, $\mu_{Y.X} = \mu_Y$ and, hence, the explained sum of the squares would be equal to 0. Therefore, ρ^2 also would be 0. In other words, none of the variability in Y would be explained by X . We would be just as well off predicting a Y -value on the basis of the mean of all the Y -values (μ_Y).

If the relationship were perfect, then the $\mu_{Y.X}$ values would be identical with the observed values. The explained variability would

equal the total variability, and ρ^2 would equal 1. That is, all the variability in Y would be accounted for by its relationship to X .

Hence, ρ^2 can range from 0 (no relationship) to 1 (perfect relationship). Its square root, the coefficient of correlation, ρ , can be either positive or negative and, hence, can range from -1 to $+1$. If the relationship of X and Y is direct—that is, as one increases the other increases—the sign of ρ is positive. If the relationship is inverse, the sign of ρ is negative. Thus, the coefficient of correlation and the slope of the regression relationship (B) have the same sign.

By convention, we refer to the degree of relationship as the coefficient of correlation (ρ). However, it is simpler to express its meaning in terms of ρ^2 , which also is known as the *coefficient of determination*.

In the above discussion we have assumed that the bivariate population was known. In practical situations, we must estimate the coefficient of correlation from a sample. In order to use statistical theory, we must make assumptions about the population. The simple linear-regression model is not a satisfactory model for this purpose. It is necessary to assume a more elaborate model which is known as the *bivariate normal distribution*.

The assumptions of this correlation model may be summarized as follows:

1. The mean of the Y -values for each given X ($\mu_{Y \cdot X}$) is on a regression line:

$$\mu_{Y \cdot X} = A + BX.$$

2. The mean of the X -values for each given Y ($\mu_{X \cdot Y}$) is on another regression line:

$$\mu_{X \cdot Y} = A' + B'Y.$$

(The two regression equations will be the same only if ρ^2 equals 1.)

3. The Y -values for every X -value are normally distributed about the regression line with constant variance ($\sigma_{Y \cdot X}^2$). The X -values for each Y -value are also normally distributed about the regression line with constant variance ($\sigma_{X \cdot Y}^2$).

In addition, this correlation model, unlike the simple linear-regression model, does not assume that the X -values are fixed (error-free). Therefore, the sampling schemes differ. In the regression model, sample Y -values are observed for various values of X —these values of X are fixed by the investigator. In the correlation model, pairs of X - and Y -values are observed on a sample of elementary units selected at random.

13.9 Sampling Theory for Correlation

In some situations a decision must be made as to whether or not association between two variables exists; that is, as to whether or not $\rho = 0$. Let us see how this decision can be made on the basis of a simple random sample of observations if the assumptions of the model described above hold.

First of all, we note that the *sample* coefficient of correlation is the pertinent statistic. It will be denoted by r , and it represents the degree of association in a sample. A simple formula for computing the value of r for a given sample is

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Let us first consider the computation of r from this formula.

Example 13.23 Two crews are making fuel tanks. The proportion of defective tanks varies from day to day for both crews. Management wishes to determine whether the same causes affect the operation of both crews. They reason that if the proportions of defectives for both crews are positively associated with each other from day to day, then common causes are most likely affecting their work. If there is no correlation, then most likely different factors are affecting the work of each crew.

Let us assume the following sample data are given. A small sample is used to simplify the arithmetic.

WORK SHEET FOR COMPUTING VALUE OF r
Percentage of Tanks with Leaks

Day of Month	Crew A (x)	Crew B (y)	xy	x^2	y^2
1.....	8	25	200	64	625
2.....	6	11	66	36	121
3.....	6	1	6	36	1
4.....	9	3	27	81	9
5.....	1	10	10	1	100
Total.....	30	50	309	218	856

$$\begin{aligned}
 r &= \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \\
 &= \frac{5(309) - 30(50)}{\sqrt{[5(218) - (30)^2][5(856) - (50)^2]}} \\
 &= \frac{45}{\sqrt{(190)(1,750)}} \\
 &= .08
 \end{aligned}$$

The value of r provides an estimate of ρ . Thus, we may write, for Example 13.23,

$$\hat{\rho} = .08.$$

This can be taken to mean that $\hat{\rho}^2$ is equal to about .0064. Since ρ^2 is the ratio of explained variation to total variation, we thus estimate that: Less than 1 per cent of the variation in the percentages of defective tanks made by crew A is associated with the percentages of defective tanks made by crew B.

Of course, since r is a random variable, the above estimate of ρ is subject to the risk of sampling error. The measurement of this error, however, is beyond our survey of the matter.

In any event, the decision problem formulated in the illustration is not one of estimation. It is one of deciding whether ρ is equal to or less than 0 against the alternative that ρ is greater than 0 (i.e., that it is positive). Certainly a value of r equal to .08 might have come from a population for which $\rho = 0$.

Let us, therefore, determine a statistical decision rule to decide the problem. We consider only the case for which the sample size is fixed. In this case only one risk, the α -risk, need be set. Let α be the risk of assuming some positive association when actually $\rho = 0$. The table below gives critical values of r for $\alpha = .05$ and $\alpha = .01$ for different sample sizes. If the value of a sample r is equal to or less than the critical value of r given in the table, we should assume that no positive relationship exists. If the sample r is greater than the critical value, we should assume that the variables are positively correlated.

CRITICAL VALUES OF THE
CORRELATION COEFFICIENT

Sample Size n	α -Risk	
	.05	.01
5	.81	.93
10	.55	.72
20	.38	.52
50	.23	.32
100	.16	.23

Let us assume that in Example 13.23 management is willing to assume an α -risk of .05. That is, they are willing to run the risk of assuming that correlation exists when actually there is none 5 times

in 100. Since $r = .08$ is less than the critical value of r (.81) for $n = 5$, they should conclude that there is no relationship between the proportion of defective fuel tanks produced by the two crews. Hence, they should conclude that different cause systems are affecting the work of the two groups, and they should act accordingly.

13.10 Rank Correlation

To apply the statistical decision rule covered in the last section properly, one must assume a bivariate normal population. If no such assumption is warranted, a rule based on an alternative statistic may be formulated. This statistic is called the *coefficient of rank correlation*. Its use requires that we reach a decision, not on the basis of the degree of association between actual observations shown in a sample but rather on the basis of the association between the *ranks* of the observations. Let us consider the following illustrative situation.

Example 13.24 It has been suggested to a sales executive that greater technical knowledge of the company's product by salesmen would increase sales and, hence, that it would be advisable to introduce a short training program for salesmen. As a test of the plausibility of this suggestion, he decides to investigate whether salesmen with greater technical knowledge actually do sell more.

Ten salesmen have been selling the product for the past 6 months. These salesmen are given a test to measure their technical knowledge. The test grades are to be compared with data on the extent to which the salesmen met, exceeded, or fell short of their sales quotas. The data are as follows:

Salesman	Test Score (Per Cent)	Percentage of Quota Met
A.....	63	128
B.....	82	75
C.....	96	150
D.....	57	85
E.....	82	115
F.....	90	110
G.....	50	100
H.....	70	104
I.....	74	100
J.....	87	90

These 10 bivariate observations will be the basis for a decision on the plausibility of the claim. A positive association would indicate that there is a degree of validity to the claim that a greater technical

knowledge by salesmen enhances sales. No relationship, or a negative relationship, would indicate a lack of validity.

If the coefficient of rank correlation is to be used as the pertinent statistic, it is first necessary to rank the observations for each variable. The actual observations given above are ranked in descending order in the work sheet that follows:

WORK SHEET FOR COMPUTATION OF COEFFICIENT
OF RANK CORRELATION

Salesman	Test Score Rank	Percentage of Quota Rank	d	d^2
A.....	8	2	6	36
B.....	4.5	10	-5.5	30.25
C.....	1	1	0	0
D.....	9	9	0	0
E.....	4.5	3	1.5	2.25
F.....	2	4	-2	4
G.....	10	6.5	3.5	12.25
H.....	7	5	2	4
I.....	6	6.5	-0.5	0.25
J.....	3	8	-5	25
Total.....			0	114.00

Observe that salesman C with the highest test score (96) has a rank of 1; salesman F with a score of 90 has a rank of 2, and so on. Note, however, that there is a tie between salesman B and E for fourth and fifth place. Each salesman is, therefore, given a rank of 4.5—the value midway between 4 and 5. The observations for the per cent of quota met are ranked in a similar fashion. Again note that salesman G and I are tied for sixth and seventh place with 100 per cent. Each salesman is given a rank of 6.5—the value midway between 6 and 7.

The value of the coefficient of rank correlation (designated as r') is given by the following formula:

$$r' = 1 - \frac{6\sum d^2}{n(n^2 - 1)},$$

where d is the difference between ranks, and n , the sample size. The values of d and d^2 for this example are computed in the work sheet given above. The value of r' for the relationship between technical knowledge and sales performance is

$$\begin{aligned} r' &= 1 - \frac{6(114)}{10(100 - 1)} \\ &= 1 - \frac{684}{990} \\ &= 1 - .69 \\ &= .31. \end{aligned}$$

The coefficient of rank correlation is a statistic and is subject to sampling error. Its value of .31 in the last example does not necessarily imply positive association in the population. A statistical decision rule to provide a basis for the decision on assuming or not assuming positive association must be formulated. To formulate such a rule, an α -risk must be fixed where α is the risk of assuming a positive association if actually there is none. The table below gives critical values of r' for $\alpha = .025$, for some sample sizes. If the value of a sample r' is equal to or less than the critical value of r' given in the table, we should assume that no positive relationship exists. If the sample r' is greater than the critical value, we should assume that the variables are positively correlated.

CRITICAL VALUES OF THE COEFFICIENT
OF RANK CORRELATION

Sample Size n	α -Risk .025
10.....	.65
15.....	.43
20.....	.37
25.....	.34
30.....	.31
50.....	.23
100.....	.16

Let us assume that an α -risk of .025 is assumed for the problem described in Example 13.24. This is the risk of assuming a positive relationship when actually no association exists. Since r' is .31 and this is less than the critical value (.65), for $n = 10$, no positive association should be assumed. The decision would be not to introduce the specialized training program at this time.

As noted earlier, the advantage of using the coefficient of rank correlation rather than r itself as the criterion for decision is that no assumption of a bivariate normal population need be made. Since this assumption of normality often cannot be made in a practical problem, rank correlation has wide applicability. However, there are disadvantages to its use. For one thing, the use of r' rather than r usually will involve greater β -risks. Secondly, the use of r' in a decision rule actually results in a test of whether or not there is any *rank* correlation in the population. Strictly speaking, it does not provide a basis for deciding whether or not $\rho = 0$. This is important because a high coefficient of rank correlation does not necessarily indicate a high coefficient of correlation. The rank meas-

ure takes only the ranks of the observations and not their magnitudes into consideration.

13.11 You Should Now Know That

Decisions by association involve decisions made on the basis of the relationship between associated variables.

There are two aspects of decisions of association: (1) regression—which relates to the nature of the relationship, and (2) correlation—which relates to the degree of the relationship.

A simple regression relationship is one between two variables—a dependent variable and one independent variable.

Multiple regression relationships involve a dependent variable and more than one independent variable.

Regression relationships also may be linear or nonlinear, depending on whether or not a straight line describes the relationship.

Scatter diagrams are useful graphic devices for investigating the nature of the relationship between two variables.

Association does not necessarily imply causation.

One useful statistical model in many regression problems is the simple linear-regression model.

There are two sources of errors in predictions based upon regression estimates: (1) errors due to an imperfect relationship, and (2) errors due to sampling.

The errors due to an imperfect relationship in the population are measured by the standard deviation of regression.

To interpret any regression estimate properly, one must consider the standard error of that estimate.

In some correlation problems, a bivariate normal population may be assumed.

In such a case, the coefficient of correlation, ρ , is a measure of the degree of relationship.

The square of this measure is called the coefficient of determination, ρ^2 , and may be interpreted as the proportion of the variability in a dependent variable which is explained by its association with the independent variable.

The sample correlation coefficient, r , is a statistic which may be used as a basis for deciding whether or not ρ is zero.

An alternative test statistic, the coefficient of rank correlation, may be used as a basis for deciding whether or not association in a popula-

tion exists even if no assumption of normality in the population is warranted.

13.12 You Should Now Be Able to Solve

1. In each of the following situations would a regression model or a correlation model be applicable? Explain each of your answers.

- A study is to be conducted to determine the relationship between gasoline consumption and the speed of an automobile.
- A study is to be conducted to determine the relationship between high school grades and college grades.
- A study is required in order to determine the relationship between the weight of a plane and the distance required for take-off.
- The relationship between education of employees and rate of turnover on a given job is to be studied.

2. In each of the following situations, discuss whether or not a cause-and-effect relationship exists between the variables. Explain your answers.

- A positive relationship between net sales/assets ratios and net sales/inventory ratios for various companies in the same industry.
- The direct relationship, over time, between the annual real estate tax rate and the total value of real estate in a community.
- The direct relationship between heights of school children and their reading abilities.
- The direct relationship between the total liabilities of a company and the price of its common stock over a period of years.

3. For each of the following sets of data draw a scatter diagram and discuss whether the relationship appears to be linear or curvilinear.

(a)

Speed of Automobile (Miles per Hour):	10	20	30	40	50	60	70	80	90
Distance Necessary to Stop after Signal (Feet):	16	42	78	124	180	246	322	408	504

(b)

Hardness of Synthetic Rubber:	45	55	61	56	66	71	86	48	62	80
Abrasion Loss of Synthetic Rubber:	322	286	251	284	246	187	130	310	241	165

(c)

Experience with Assembly Operation (Days):	5	5	5	10	10	10	15	15	15
Time Taken to Perform an Assembly (Minutes):	36.2	41.8	35.7	22.8	19.3	24.9	17.0	17.5	16.8

4. A home laundry service desires to determine the relationship between age of trucks and annual repair costs. It will use the relationship to estimate repair costs, and to determine a replacement policy for its fleet of trucks. The following data are assembled:

Truck	Age (Years)	Annual Repair Cost (in Dollars)
1.....	1	22
2.....	1	75
3.....	1	104
4.....	1	0
5.....	1	0
6.....	2	37
7.....	2	69
8.....	2	81
9.....	2	148
10.....	2	46
11.....	3	105
12.....	3	102
13.....	3	188
14.....	3	65
15.....	3	36
16.....	4	108
17.....	4	63
18.....	4	235
19.....	4	214
20.....	4	127
21.....	5	132
22.....	5	117
23.....	5	174
24.....	5	127
25.....	5	153

- Plot a scatter diagram for these data.
- Estimate a regression line to describe the relationship between annual repair cost and age of truck.
- Discuss the extent to which the assumptions of the simple linear-regression model are valid for this problem.
- Should it be concluded that a direct linear relationship exists between the variables? Explain your reasoning.
- Would you recommend the relationship as a means of estimating repair costs? Explain.

5. Many states employ the assessed valuation of real property as a criterion of the ability of units of local government to finance their own operations. As a check on the adequacy of this criterion, one state wishes to compare the per capita assessed valuation of gov-

ernmental units with their median family income. A sample of 25 districts is selected, and the results are as follows:

District	Per Capita Assessed Value (in Dollars)	Median Family Income (in Dollars)
1.....	2,600	3,570
2.....	2,810	3,680
3.....	2,820	4,420
4.....	2,800	4,070
5.....	2,540	3,870
6.....	3,020	3,940
7.....	2,870	4,290
8.....	2,710	4,040
9.....	2,680	3,550
10.....	2,650	3,320
11.....	2,540	3,560
12.....	2,610	3,510
13.....	2,760	3,900
14.....	2,820	4,320
15.....	2,630	3,790
16.....	2,460	3,570
17.....	2,920	4,350
18.....	2,610	3,920
19.....	2,600	3,790
20.....	2,480	3,510
21.....	2,860	4,050
22.....	3,150	4,520
23.....	2,440	3,580
24.....	2,930	3,420
25.....	2,620	3,600

- a) Estimate the population coefficient of correlation and interpret your result.
 - b) Discuss the extent to which the assumptions of the bivariate normal population model are valid for this problem.
 - c) Should it be concluded that some positive association exists between the variables? Explain your reasoning.
6. For the problem situation given in Question 5, formulate a statistical decision rule in terms of the coefficient of rank correlation. Apply the decision rule. Why do the values of the coefficient of correlation and the coefficient of rank correlation computed for this and the last question differ?
7. The following are data for disposable personal income and consumer expenditures for furniture from 1947 through 1957:

Year	Expenditures for Furniture (Billions of Dollars)	Disposable Personal In- come (Billions of Dollars)
1947.....	2.55	170
1948.....	2.78	189
1949.....	2.69	190
1950.....	3.07	208
1951.....	3.18	228
1952.....	3.39	239
1953.....	3.58	252
1954.....	3.64	257
1955.....	4.18	274
1956.....	4.43	290
1957.....	4.34	305

- Plot the scatter diagram for the data.
- Are the assumptions of the simple linear-regression model satisfied in this problem? Explain why or why not.
- Determine a regression equation.
- How can you determine the accuracy of this relationship?

13.13 You Will Also Find That

The topics which we have considered in this chapter—including correlation and regression in general, the simple linear-regression model, standard errors, the coefficient of correlation, and so forth—also are discussed in the following textbooks:

HIRSCH, WERNER Z. *Introduction to Modern Statistics*, pp. 247–80. New York: Macmillan Co., 1957.

MCCARTHY, PHILLIP J. *Introduction to Statistical Reasoning*, pp. 332–88. New York: McGraw-Hill Book Co., Inc., 1957.

SPROWLS, R. CLAY. *Elementary Statistics*, pp. 222–59. New York: McGraw-Hill Book Co., Inc., 1955.

WALLIS, W. ALLEN, and ROBERTS, HARRY V. *Statistics: A New Approach*, pp. 524–56. Glencoe, Ill.: The Free Press, 1956.

Forecasting with Time Series

14.1 Some Remarks on Forecasting

In this chapter we shall concern ourselves with some aspects of business forecasting. Forecasting problems arise when decisions must be made concerning future populations. But since future populations cannot be studied directly, forecasting involves the methods and theory of making decisions for the future on the basis of a study of past and present populations.

We already have considered such problems in previous chapters. Thus, in quality control, the pattern of process variability in the future is predicted on the assumption that a process remains in control. Then, again, in decisions by association, predictions are made about the future behavior of a variable on the basis of its present relationship with an associated variable.

In each of these instances predictions are statistical forecasts. However, it is impossible to forecast the future precisely—there always must be some range of error allowed for in the forecast. Statistical forecasts are those for which we can use the mathematical theory of probability to measure the risks of errors in predictions. To apply statistical theory for prediction, however, we had to make certain assumptions. To use the quality control chart for purposes of prediction, we assumed that the process was a random process—operating in statistical control. In order to use the regression line to predict, we had to assume that the population satisfied the regression model.

There are many forecasting problems that businessmen face for which the populations of interest do not satisfy the conditions necessary for the application of statistical theory. In these situations, al-

though the techniques used may be the same, the risks of an incorrect forecast must be evaluated on the basis of expert judgment. The decision procedure then is only partly statistical. We already have illustrated this point with reference to the judgment applications of regression relationships.

Unfortunately, forecasting is a necessary evil. It involves future planning by businessmen who must continually determine their needs and demands for tomorrow, next week, and next year. A business may face decisions of whether or not to increase production during the second half, of what sales quotas should be set for the first quarter, and so on. To ignore the existence of these problems and to refuse to forecast is a rather naïve approach. "No forecasting" is, in itself, actually a forecast that business conditions will not change.

The remainder of this chapter will be devoted to a discussion of *time series analysis* and its use in forecasting. Unfortunately, there is as yet no complete body of statistical theory developed for this subject. Therefore, the application of the methods to be discussed to forecasting problems will depend to a large extent on expert judgment. However, the statistician by pointing out the limitations of these methods can be of assistance to the expert in the substantive field where such techniques will be applied. The methods reduce the area within which judgment must operate and thus sharpen the analysis.

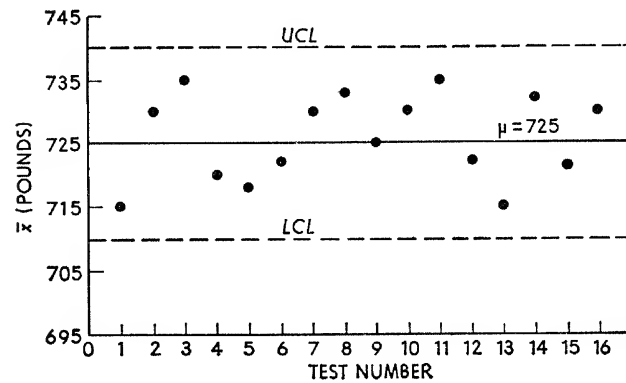
14.2 Time Series and Forecasting

A time series consists of a set of observations made at different periods of time. We already have encountered time series in our study of control charts. A control chart presents sample means, ranges, proportions, or standard deviations observed at uniform intervals of time.

Example 14.1 Two control charts are illustrated on the following page. The first chart (a) is the control chart for the average shearing strength of spot welds. This chart is similar to those presented in Chapter 11.

Chart (b) is a control chart for the average length of steel bars. This control chart presents a time series which differs from those we have considered heretofore. It represents a condition in which the tool used to cut the bars is subject to rapid wear.

a) CONTROL CHART FOR MEAN SHEARING STRENGTH OF SPOT WELDS



b) CONTROL CHART FOR MEAN LENGTH OF STEEL BARS

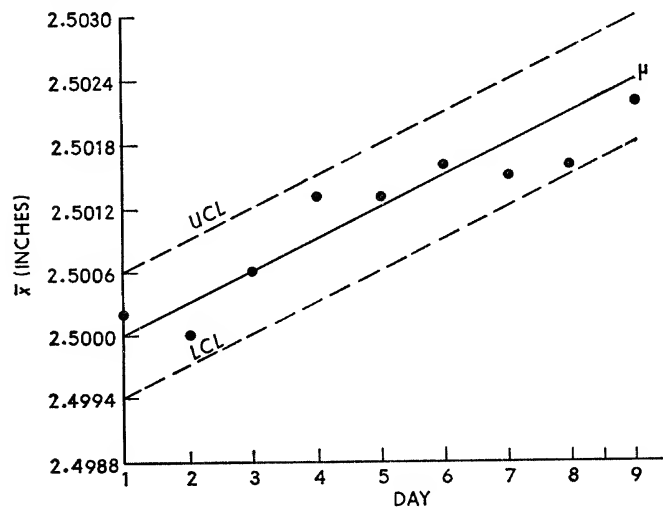


Chart (a) in the preceding example presents a very simple time series. The population (process) mean remains constant from one instant of time to the next. The individual sample means are distributed randomly about the population mean. The deviations about the population mean are independent of each other. That is, our knowledge of the fact that one sample mean is above the population mean in no way enables us to predict whether the succeeding sample means will be above or below the population mean.

Chart (b) represents a more complicated time series. Here, because of tool wear, the population average is shifting from one day

to the next. However, the population mean shifts in a systematic manner—3 ten-thousandths of an inch per day. In other words, the shift in the population mean is associated with time, and the line of regression of the population mean on time has a slope of .0003 inches per day. Thus, part of the upward drift in the sample averages is explained by tool wear from day to day. The difference between the sample mean observed and the population mean for that day, however, is a random event, and again the fact that the sample for one day will be above or below the population mean gives us no clue as to the nature of the deviation for succeeding or preceding days. We can say that the time series in chart (b) is composed of two elements: (1) a systematic element represented by the regression line showing the day-to-day change in the population average that is explained by tool wear, and (2) a random element due to a host of chance factors.

Most time series we encounter in business data, outside the field of quality control, are much more complicated than the time series illustrated in Example 14.1. Furthermore, in most time series the deviations of the observations about the population means are not independent but related to each other. When the values of the observations that follow each other or are near each other are correlated, we refer to this type of correlation as *serial correlation*. The presence of serial correlation makes it impossible to use the methods described earlier in this text in order to forecast the pattern of future behavior of a series on the basis of its past behavior.

Methods beyond the scope of this book do exist for estimating the degree of serial correlation in time series data. Similarly it is possible to measure the standard error of an estimate derived from data that are serially correlated, if the degree of correlation is known. Much work is being done currently on the subject of time series in the field of theoretical statistics. To date, no body of techniques that can be widely applied to practical business problems has been developed. One of the great obstacles has been the fact that the newly devised techniques require long series of computations. Perhaps the development of electronic computers will make such techniques feasible in the future. Until then, it will be necessary to put heavy reliance on judgment to forecast the pattern of behavior of future events.

14.3 Graphic Presentation of Time Series

In working with time series it is desirable to present the data graphically. Such presentations assist our analysis of time series

much as a scatter diagram did in the case of a regression problem. A graph simplifies the study of the movements of the data.

In visualizing changes in time series we are interested in two types of comparisons: *absolute* changes and *relative* changes.

Example 14.2 Consider the following figures of consumer installment credit:

1945—\$ 5.5 billion
1950—\$14.5 billion
1955—\$27.8 billion

During the 5-year period 1945–50, consumer installment credit increased \$9.0 billion. During the 5-year period 1950–55, the increase was \$13.3 billion. Thus, the absolute increase (in dollars) was greater during the second 5-year period than during the first.

The relative increase between 1945 and 1950 was approximately 164 per cent. Between 1950 and 1955 the relative change was approximately 92 per cent. Thus, although the absolute change was larger during the second 5-year period, the relative change was smaller.

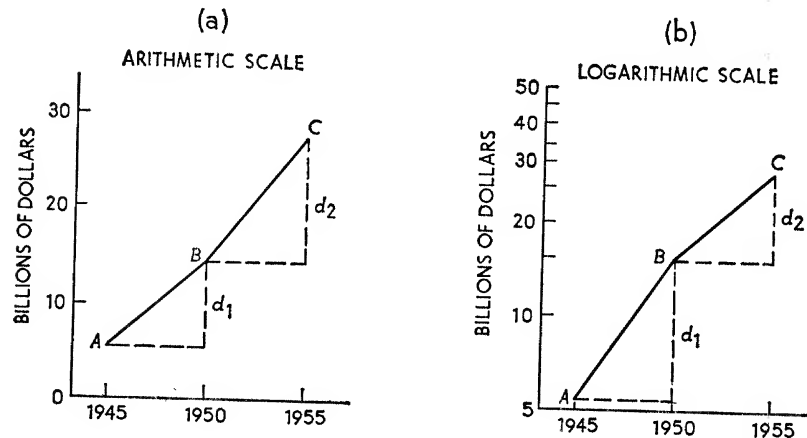
To get a visual impression of absolute changes in data, uniformly spaced or arithmetically scaled graph paper is used. We could not determine relative changes from time period to time period from such charts (without resorting to computations). Relative changes, however, can be portrayed visually on *semilogarithmic* charts (although only the absolute values of the data need to be plotted). These charts are ruled with an arithmetic scale on the X -axis and a logarithmic scale on the Y -axis—hence, the term “semilogarithmic.” *A logarithmic scale is ruled in such a manner that when we plot a value, we actually are plotting its logarithm.* The scale is ruled in the same way as the scales used for multiplication and division on a slide rule. Therefore, when data are plotted on a logarithmic scale, equal vertical distances represent equal relative changes rather than equal absolute changes.

Example 14.3 Figure 14.1 presents the data of consumer installment credit according to both arithmetic and logarithmic scales.

On the arithmetic scale chart the vertical distance (d_1) between the two points A and B is smaller than the vertical distance (d_2) between points B and C . This is so because the absolute change between 1945 and 1950 is smaller than the absolute change between 1950 and 1955. Arithmetic scale graphs picture absolute changes more clearly.

On the semilogarithmic chart the vertical distance (d_1) between A and B is greater than the vertical distance (d_2) between B and C .

Figure 14.1

CONSUMER INSTALLMENT CREDIT IN THE UNITED STATES
1945 1950 1955

This reflects the fact that the relative change between 1945 and 1950 is greater than the relative change between 1950 and 1955. Semilogarithmic graphs depict relative changes more clearly.

Logarithmic scales are ruled as in Figure 14.2. The scale may be single tiered as in (a), double tiered as in (c), or multiple tiered. Since equal vertical distances represent equal ratios, the distance from 1 to 2 is the same as the distance from 2 to 4, and from 3 to 6, and so on. For the same reason the distance from 1 to 2 (ratio 2:1) is greater than the distance from 2 to 3 (ratio 3:2 or 1.5 to 1). The greater the relative change, the greater is the vertical distance.

Example 14.4 Figure 14.2 presents five illustrations of logarithmic scales. Scale (a) illustrates the manner in which most types of commercially prepared logarithmic scales appear when purchased. Scale (b) is prepared for data which range from \$10 to \$100. You will notice that since number 1 on the scale is set at \$10, number 2 must be designated as \$20. Thus, 3 is \$30, and so on.

Scale (c) is arranged for a two-tiered semilogarithmic chart of data that vary between 10 and 1,000 tons. Scale (d) is prepared for data that range from 700 to 7,000 employees. Scale (e) may be used for data which range from \$.50 to \$50.

You will observe that on each of these scales equal vertical distances represent equal ratios.

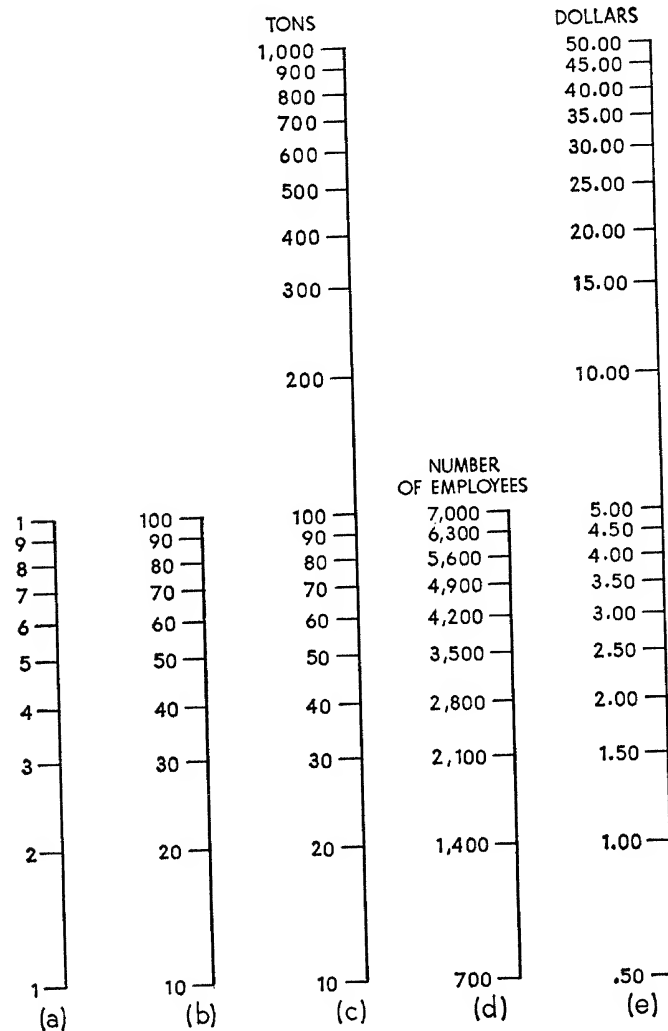
Semilogarithmic charts also are used at times to present data which vary widely in absolute amounts. At other times we may desire to present two types of data where the magnitudes of one series

vary widely from those of the second. Here again a logarithmic chart facilitates graphic comparisons.

Where semilogarithmic charts are used in reports, care must be

Figure 14.2

ILLUSTRATIONS OF LOGARITHMIC SCALES



taken to explain that equal vertical distances on the chart represent equal ratios, or equal relative changes. Otherwise, such charts may be misinterpreted by readers who are not acquainted with this form of graphic presentation.

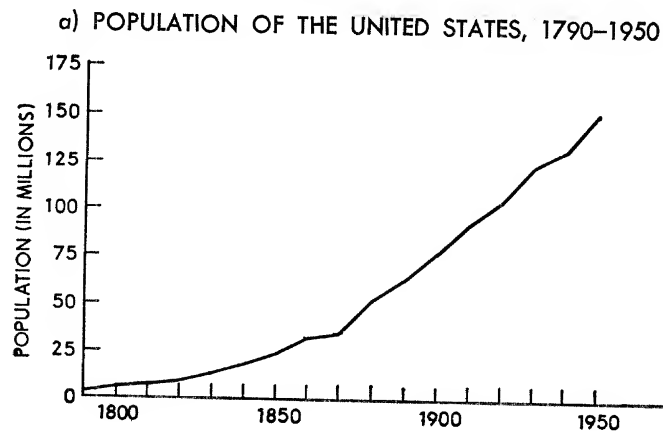
14.4 Nature of Economic Time Series

The control chart for mean length of steel bars in Example 14.1 presents a time series composed of a systematic element and a random element. The systematic element causes the shift in the population mean from one period to the next, while the random element accounts for the variability about that mean. Economic time series also can be described in such terms. However, the systematic element in economic time series is much more complicated than in the quality control illustration. There, the systematic element reflected the influence of a single factor—tool wear. Economists identify three types of systematic elements that influence the behavior of economic time series: (1) trend, (2) seasonal, and (3) cyclical. Each of these factors is the result of different forces affecting the data. In addition to the three systematic elements, there is a random element.

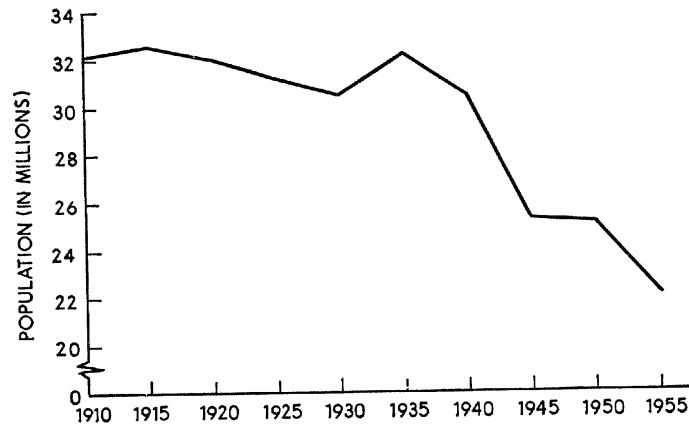
The *trend* of a time series is the resultant of the long-time, slowly moving forces affecting the data. Such forces include population growth, technological changes, changes in habits and customs, and so on. Trends can be described qualitatively as moving upward, downward, or horizontally (no trend).

Figure 14.3 includes three economic time series. You will observe that each series reflects a gradual long-term movement—trend. Thus, while the growth of the population of the United States reflects a continuous upward trend, farm population reflects a downward trend. The index of industrial production also reflects an upward trend.

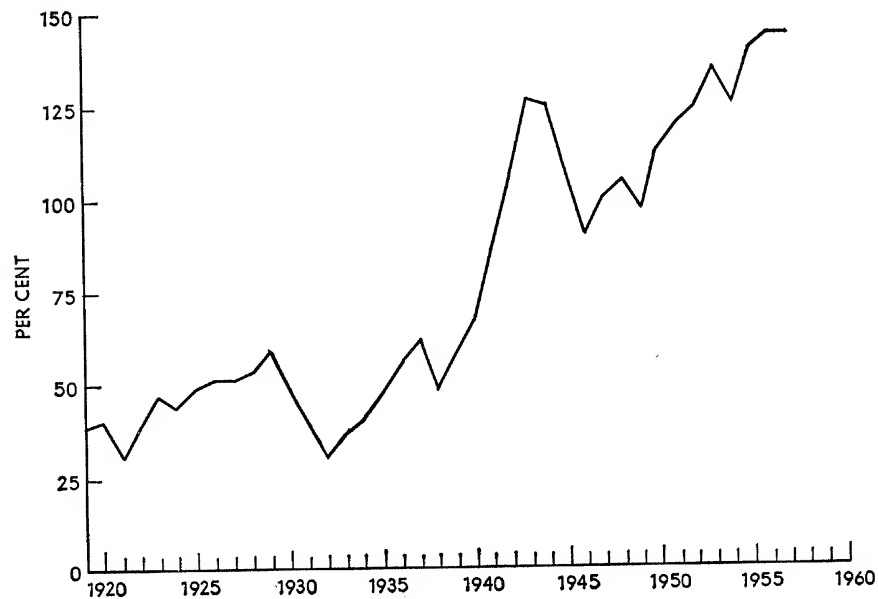
Figure 14.3



b) FARM POPULATION IN THE UNITED STATES, 1910-55



Charts (b) and (c) indicate that in addition to the long-term trend, there are oscillatory movements. Thus, farm population hit a low in 1930, increased in 1935, and then continued to decrease again. The increase in farm population during the thirties resulted from the lack of other opportunities for employment because of depressed

c) INDEX OF INDUSTRIAL PRODUCTION, 1919-57
(1947-49 = 100)

business conditions. The oscillatory movements from peaks to troughs and back to peaks in the *Index of Industrial Production* also reflect the influences of general business activity. Thus, the business recessions of 1921, 1924, 1927, 1932, 1938, 1949, 1954, and 1957 are clearly indicated. Such oscillatory movements, of more than one year in duration, that are related to ups and downs in general business are designated as *cycles*. You will observe that not all up and down movements are cycles. Thus, the peak in 1944 and the trough in 1946 are the result of the sharp increase in industrial production due to the war and the subsequent return to peacetime conditions.

Figure 14.4

INDEX OF DEPARTMENT STORE SALES IN THE UNITED STATES, 1955-58
(1947-49 = 100)

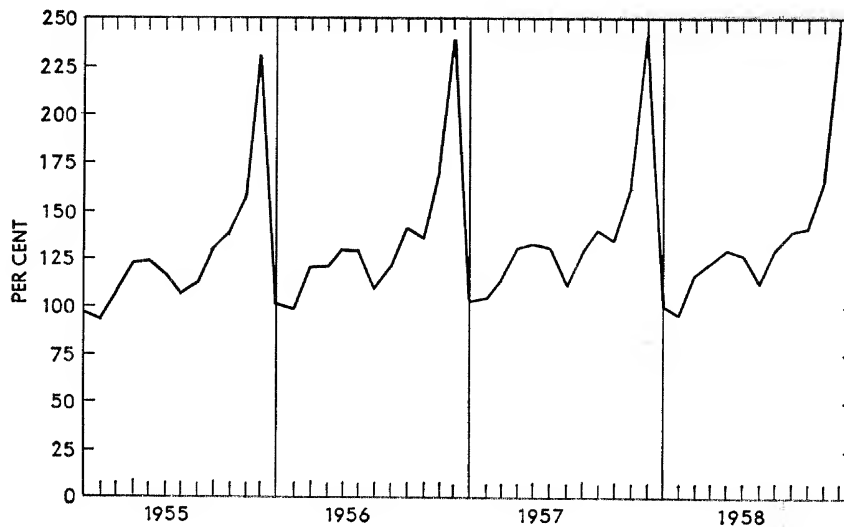


Figure 14.4 presents department store sales for the period 1955-58. The trend of this time series is slightly upward. There also is a cyclical peak in 1956 and a trough in 1957. The trend and cyclical movements, however, are obscured by the sharp up and down swings during the course of each year. This third movement in the data occurs regularly every year. For example, department store sales fall to a low at the beginning and middle of each year, and reach a peak during Easter, and a much higher peak at the end of the year. Such variations, which occur regularly during the period of a year, are known as *seasonal fluctuations*. These movements may result from

natural causes such as climatic conditions. Thus, ice cream sales increase during the summer, as does building construction. However, natural causes do not account for all seasonal fluctuations—some causes are man-made. These include (1) the custom of giving gifts at Christmas, Mother's Day, and Father's Day; (2) the change of style in clothing in the spring or fall; (3) the desire to purchase new cars in the spring; (4) the change in the date of the automobile show to counteract this desire; and so on.

Thus, Figure 14.4 portrays a monthly seasonal movement. However, you will observe that if we were to plot annual data, the monthly seasonal fluctuations would be obliterated. Seasonal movements also can be quarterly, weekly, daily, or hourly. Thus, if we had weekly data of department store sales, we might discover a regular pattern of fluctuation during the month. Such movements are obscured, however, when monthly data are used. If daily data were available, regular fluctuations within a week could be observed. Thus, commercial banks transactions reach a peak on the day when firms in the area pay their employees. If hourly data were available, then the movements within the course of a day could be observed. For example, such hourly seasonal movements are important to restaurants.

Finally, in each of these charts there are certain movements which can be classified as random movements. That is, they result from a host of chance forces.

Time series analysis concerns itself with identification of the systematic elements. On the basis of past trends, past seasonal variations, and past cyclical movements, attempts are made to forecast the future course of events. The existence of serial correlation, however, makes it difficult to identify the systematic movements. Thus, a random event might result in a wide deviation from the population average. Because of serial correlation, the succeeding values would move back slowly to the population mean. Thus, if the set of observations were analyzed soon after the occurrence of a random event, the random movement might be misinterpreted as an upward trend. The same effect might also create the illusion of a cyclical movement when observed at a later period. By that time, the series will have returned to the population average or even have dropped below it. We thus would observe as a cyclical peak, a movement in the data that was a combination of a random movement and serial correlation of the deviation of the series from its population average.

In analyzing a time series for its systematic and random components, the assumption is made that the four elements represent the result of different factors. The isolation of the three systematic elements will, therefore, throw some light on the nature of the factors affecting them. This knowledge is then applied to project the effect of the systematic elements into the future. After allowances for random effects a forecast can be made on the basis of judgment.

Thus, if Y_t is used to symbolize the actual observation at time t , we shall express Y_t in terms of the series' trend at time t (T_t), the seasonal influence at time t (S_t), the cyclical influence (C_t), and random factors (R_t). One relationship often assumed is that

$$Y_t = T_t \times S_t \times C_t \times R_t.$$

This equation states that the actual value of a variable at any time, Y_t , is the product of the four elements considered. It implies that the magnitudes of T , S , C , and R are interrelated. This is not the only relationship we could assume, though it is the one most generally used. For example, an alternative relationship is

$$Y_t = T_t + S_t + C_t + R_t.$$

This implies that the magnitudes of the four elements are additive. In other words, the magnitude of one element would not be affected by the magnitude of any other.

As noted, time series analysis deals with the isolation of the T , S , and C factors from the original data. In subsequent sections, then, we shall discuss some of the techniques that may be used to isolate these elements.

14.5 The Analysis of the Trend Element

As we have stated, trend represents the resultant of the long-time, slowly moving forces that affect the data. In order to determine the nature of the trend, a first step is to plot the series on a chart. The chart, as we have stated previously, then serves a purpose similar to the scatter diagram in a regression problem. In fact, a trend relationship may be looked upon as a regression relationship with time as the independent variable.

Before attempting any analysis of time series data, it is essential to see that the data are comparable for the entire period. Data gathered at different times are often based on units and terms differently defined, or are observations from different populations. In

such instances the data must be adjusted. Otherwise, differences in definition or coverage might be interpreted as evidence of systematic movements.

Example 14.5 Consider the series on the number of clerical and kindred workers employed in the United States during the census years from 1900 to 1950. There have been changes in the definition of the term "employed" from census to census. There also have been changes in the coverage of the "clerical and kindred workers" classification. Thus, for example, the census of 1950 excluded accountants and auditors from this classification, while prior censuses included them. It is necessary, therefore, to adjust the data before subjecting it to time series analysis.

It is not always possible to make such adjustments. However, when the changes in definition or coverage are not large enough to affect the analysis of the series, we may use the unadjusted data. When the changes are more significant, it is necessary to break the data down into time periods of uniform coverage and definitions and then analyze each grouping independently.

A Statistical Test for Trend. Even though it seems that there is an upward or downward drift in the data, we must remember that such movements could represent the fluctuations of a random process. An upward or downward movement could occur even though the process had no systematic element and actually was behaving in a completely random manner. Therefore, before a trend is fitted, we should decide whether, on the basis of the evidence at hand, the series represents a sample from a completely random process or reflects a systematic movement.

You should recognize that we are faced with a statistical decision problem where two alternate courses of action are possible:

1. No trend should be fitted since the series represents a random process.
2. A trend should be fitted since the series indicates the presence of systematic forces.

Consider the differences between successive observations in a time series. The number of differences always will be one less than the number of observations. If many samples were drawn from a random process, then the expected *number* of positive differences between successive observations would be equal to the expected number of negative differences. If we call the expected number of differences having a positive sign, $E(s)$, we have

$$E(s) = \frac{n-1}{2},$$

where n is the number of observations. Statisticians have developed a formula for the standard error of s under the assumption of a completely random process. The formula is

$$\sigma_s = \sqrt{\frac{n+1}{12}}.$$

For large samples the sampling distribution of the number of positive differences in a sample is approximated by the normal curve. The normal deviate is

$$z = \frac{s - E(s)}{\sigma_s},$$

where s is the observed number of positive differences in a sample of n observations.

Since n usually is given in a time series problem, in order to determine a decision rule we need set only the α -risk (i.e., the probability of fitting a trend to the data when actually no trend exists). Let's look at an application.

Example 14.6 Data for the Federal Reserve Board Index of Industrial Production from 1919 to 1957 are given in Table 14.1. These data already have been presented in Figure 14.3 (c). The chart in this figure seems to indicate a systematic upward movement. We now would like to apply the test discussed above to determine whether or not the appearance of an upward movement could have resulted from a completely random process.

Table 14.1

INDEX OF INDUSTRIAL PRODUCTION FOR THE UNITED STATES, 1919-57
(1947-49 Average = 100)

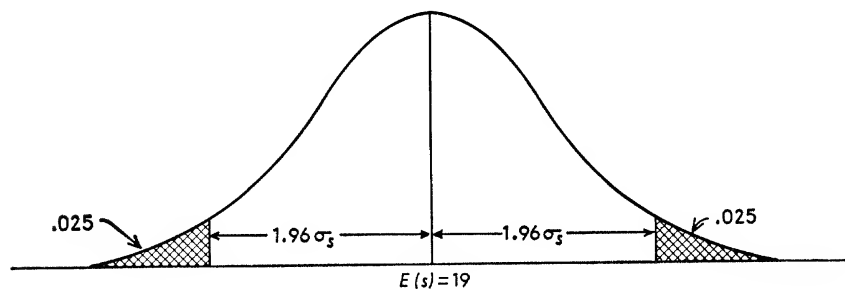
Year	Index Number	Sign of First Difference	Year	Index Number	Sign of First Difference	Year	Index Number	Sign of First Difference
1919.....	39		1932....	31	-	1945....	107	-
1920.....	41	+	1933....	37	+	1946....	90	-
1921.....	31	-	1934....	40	+	1947....	100	+
1922.....	39	+	1935....	47	+	1948....	104	+
1923.....	47	+	1936....	56	+	1949....	97	-
1924.....	44	-	1937....	61	+	1950....	112	+
1925.....	49	+	1938....	48	-	1951....	120	+
1926.....	51	+	1939....	58	+	1952....	124	+
1927.....	51	*	1940....	67	+	1953....	134	+
1928.....	53	+	1941....	87	+	1954....	125	-
1929.....	59	+	1942....	106	+	1955....	139	+
1930.....	49	-	1943....	127	+	1956....	143	+
1931.....	40	-	1944....	125	-	1957....	143	*

* There was no change between 1926 and 1927 nor between 1956 and 1957. These are counted as one half of a positive difference.

Let us first determine a statistical decision rule.

$$\begin{aligned}
 E(s) &= \frac{n-1}{2} \\
 &= \frac{39-1}{2} \\
 &= 19 \\
 \sigma_s &= \sqrt{\frac{n+1}{12}} \\
 &= \sqrt{\frac{40}{12}} \\
 &= \sqrt{3.33} \\
 &= 1.8
 \end{aligned}$$

Hence, the sampling distribution of s in repeated sampling from a purely random process would appear as in the diagram below. Let us assume an α -risk of .05—the shaded area below:



The appropriate decision rule is thus: If the sample of 39 observations results in a normal deviate of less than 1.96, no trend need be fitted. If the normal deviate is 1.96 or greater, a trend should be fitted.

If we count the number of positive differences shown in Table 14.1, we find $s = 25$. Hence,

$$\begin{aligned}
 z &= \frac{s - E(s)}{\sigma_s} \\
 &= \frac{25 - 19}{1.8} \\
 &= 3.3
 \end{aligned}$$

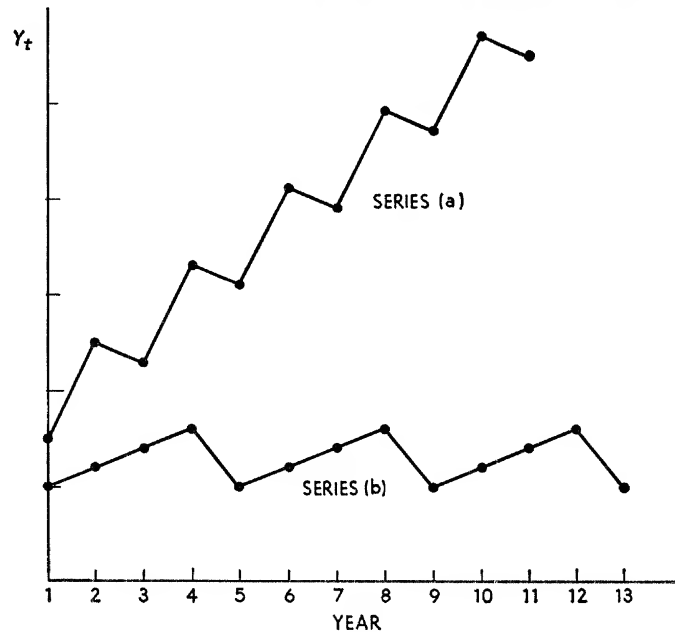
Since z is greater than 1.96, we conclude that a trend should be fitted.

A few words of caution concerning this test are in order. The test considers only the direction of movement from time period to time period and not the magnitude of such movements. Thus, the magnitude of the upward movements might be large, while the downward,

although equal in number, might be very small. The presence of trend in such instances would not be revealed by this test. Similarly, this test will indicate the presence of a trend where no trend exists if many small upward movements are canceled out by a few sharp downward movements. However, in such instances the chart of the data should reveal whether situations like those just described exist. We then would know that the use of this test for the presence of trend would be misleading. Figure 14.5 illustrates time series for

Figure 14.5

ILLUSTRATIONS OF HYPOTHETICAL TIME SERIES WHERE FIRST DIFFERENCE TEST FOR TREND DOES NOT APPLY



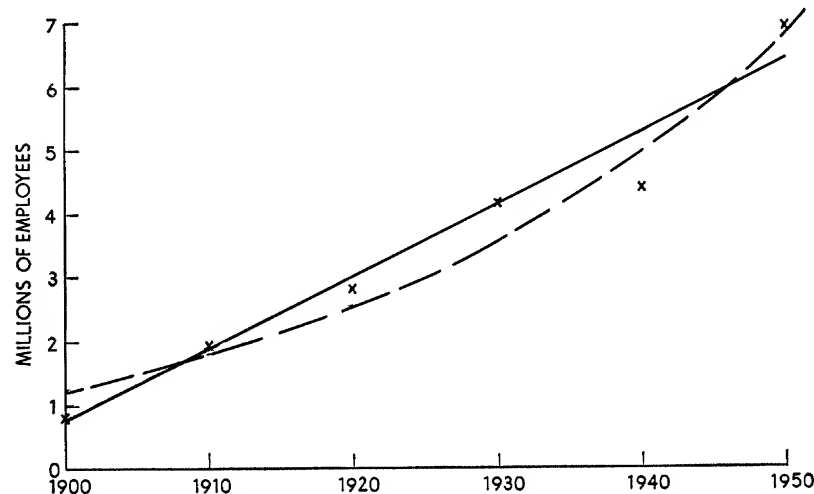
which: (a) the above test would indicate a random process even though a trend were present, and (b) the test would indicate the existence of trend even though the process were completely random.

Fitting the Trend. Once the presence of a systematic trend is indicated by a statistical test or on the basis of judgment, we must examine the data to develop a theory concerning the trend. Is the trend represented by a straight line or a curve? If a curve, what is the exact nature of the curve? To get a better idea of the considerations involved in deciding upon the nature of the trend, the following problem is instructive.

Example 14.7 Figure 14.6 illustrates the growth of the number of clerical and kindred workers employed in the United States from 1900 to 1950, by 10-year periods. The data are adjusted for differences in definition and coverage.

Figure 14.6

CLERICAL AND KINDRED WORKERS EMPLOYED IN THE UNITED STATES
FOR CENSUS YEARS 1900-1950



Observe that two possible trend lines are drawn to illustrate the long-term growth of the data. The curve weaves in and out of the actual observations. These deviations could be looked upon as random effects. Hence, if we decide upon the curve rather than the straight line, our rationale is that the data represent trend and random elements.

However, if we decide upon the straight line trend, we imply that forces other than those giving rise to trend and random elements are operative. The fact that the point for 1940 is far below the straight line trend would be attributed to the presence of cyclical forces affecting the data. The census results for 1940 would be taken to reflect the impact of the decline in business activity during the 1930's. Although the number of clerical workers increased during this period, our choice would imply that the increase was retarded because of the depressed conditions of that decade.

The brief discussion in the above example should enable you to see the subjective nature of methods for determining the trend. The choice of an appropriate trend line must be made on the basis of knowledge of the data involved and on the basis of economic theory. It is not always a simple matter to determine the nature of a trend. In fact, two experts might devise different, though logical, theories as to the nature of the trend of a given time series.

Once the form of a trend line is decided upon, it can be fitted to

the data by either freehand methods or by mathematically derived formulas. The advantage in mathematically fitted trends is that such trends can be extrapolated more easily. A freehand curve can be projected in many ways. It often is claimed that mathematical methods have the advantage that all persons fitting the trend by this method will get the same results. This claim is of doubtful validity. In the final analysis a subjective decision on the adequacy of a trend must be made on the degree to which the fitted curve satisfies the concept of trend. It, therefore, would appear to be much simpler to draw freehand the curve that seems best to meet our theory of trend. There is no advantage in the fact that all who fit a mathematical curve will get the same results if they all get the same incorrect results. Furthermore, in fitting a mathematical curve there always is the danger of ascribing to the curve a rigor and precision far beyond the subjective analysis upon which its choice is based. This would be similar to measuring the dimensions of a dining-room table to the nearest thousandth of an inch in order to purchase a table cloth.

However, if a mathematical curve is to be used, we can fit a least-squares straight-line trend with the same formulas as those applied in fitting a regression line. In time series analysis the X -variable represents time. Hence, our straight-line equation is

$$T_t = a + bX,$$

where T_t is the trend of the time series. Let us see how we may estimate the trend line for a given series.

Example 14.8 The straight-line trend in Figure 14.6 is drawn in freehand. The fitting of a least-squares straight line to the same data is illustrated below. Note that time is the X -variable. X -values have been assigned to each year as 0, 1, . . . , for convenience. Hence, we must remember that $X = 0$ in 1900 and that for each 10 years, X increases one unit.

COMPUTATION OF LEAST-SQUARES TREND LINE FOR EMPLOYED
CLERICAL AND KINDRED WORKERS, 1900-1950

Year	X	Workers (Millions) Y	XY	X^2
1900.....	0	0.8	0	0
1910.....	1	1.9	1.9	1
1920.....	2	2.8	5.6	4
1930.....	3	4.1	12.3	9
1940.....	4	4.4	17.6	16
1950.....	5	6.9	34.5	25
Total.....	15	20.9	71.9	55

The values of b and a are calculated by the least-squares formulas which were presented in the preceding chapter.

$$\begin{aligned}
 b &= \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \\
 &= \frac{6(71.9) - 15(20.9)}{6(55) - (15)^2} \\
 &= \frac{431.4 - 313.5}{330 - 225} \\
 &= 1.12 \\
 a &= \bar{Y} - b\bar{X} \\
 &= \frac{20.9}{6} - 1.12\left(\frac{15}{6}\right) \\
 &= 0.68
 \end{aligned}$$

Hence,

$$T_t = 0.68 + 1.12X.$$

(Origin: 1900, $X = 10$ years.)

Whenever the equation of a trend line is given, the *origin*—the year at which X is zero—and the unit of time in which X is expressed must be indicated.

In fitting any type of trend line to data marked by cyclical influences, care must be taken if the time series starts at one phase of the cycle and ends at another phase. Thus, a trend fitted to a series that starts at a low point in the cycle and ends at a high point will have an exaggerated upward slope. Similarly, a downward slope will be exaggerated, or an upward slope understated, if we start at a peak period and end at a low period. If we were to fit a trend line for the series on clerical and kindred workers from 1940 to 1950, we would have too steep a trend, unless we took into consideration the fact that the year 1940 represented a cyclical low point. Similarly, the slope of a line would not be steep enough if we were to fit the trend to the period 1910–40 and again neglected the fact that 1940 represents a low point in the cycle. This again emphasizes the importance of a thorough knowledge of the subject matter in fitting trend lines.

Similar methods are available for fitting curvilinear trends. However, we shall not deal with those methods in this text.

Projection of Trends. Trend lines may be fitted for historical reasons, i.e., to indicate the growth of a series in the past. However, their main purpose in business problems is projection. Thus, an electric power company that provides services to residents of a certain area must project the demand for power at least 5 years into the future. It takes that long to expand the capacity of such a utility.

Similarly, an oil company might project the demand for oil 25 years into the future to determine the sufficiency of available resources and to plan for future resource development. In addition, a study is now under way to determine the characteristics of the New York Metropolitan Region in 1975. Such projections are needed as a basis for intelligent regional planning and development.

It takes only a second to project a trend with ruler and pencil even for 50 years into the future. Such projections, however, rest on assumptions about the future that must be recognized.

Example 14.9 It would be a simple matter to project a freehand trend line, or a least-squares line, to predict the number of employed clerical and kindred workers for any future year. In so doing, however, we would be assuming that the impact of the trend factors influencing the growth in the past will have the same combined effect in the future. Among the more important of these forces are: population growth, growth in the labor force, average hours worked, productivity (automation, etc.), changes in relative importance of service industries, changes in size of businesses, changes in business location.

Let us discuss briefly some of the matters that would have to be considered in order to determine whether or not the trend forces would act in the future as in the past.

The rate of population growth which had been decreasing from decade to decade since 1870, had increased between 1940 and 1950. This, by itself, would tend to make for an increased rate of growth in the labor force. The trend line, however, was determined on the basis of 1900-1950 experience. Therefore, a simple extension of a trend line might underestimate the impact of the accelerated growth in population and labor force even though the proportion of the labor force going into clerical operations remained constant.

If productivity should increase in clerical professions due to the increased use of electronic equipment and other business machines, fewer people might be employed in clerical occupations. But, if the average hours of work per week should decrease in the future, this would in part counteract the results of increased productivity. Furthermore, electronic equipment is still relatively expensive, and hence its installation is economically feasible for only large firms. There has been a tendency, however, for industry to concentrate in "industrial parks." This has given rise to firms whose function is to provide clerical services for these industries with the aid of electronic equipment. Should this procedure become more and more accepted, it would allow small firms to avail themselves of increased productivity due to electronic equipment and make for a decrease in the number of clerical workers needed.

Service industries are becoming relatively more important in our economy. Generally, clerical workers comprise a large proportion of total employment in these industries. Therefore, if this tendency con-

tinues, it would tend to make for relatively greater employment of clerical workers.

We could continue in this vein, discussing other factors that would make for more rapid growth, slower growth, constant growth, or a decline in the number of employed clerical workers in the future. Expert judgment must be used to evaluate the net effect of such developments in determining the necessity for adjustment in any trend projection.

By carefully spelling out the assumptions upon which a forecast is made, a method of analysis for future events is provided. As the strength or weakness of the initial assumptions becomes apparent, adjustments can be made in the prediction. Good forecasting does not imply merely a precise estimate. Such precision often may be due to fortuitous circumstances. Good forecasting implies a system of continuous analysis—a system that will disclose why initial forecasts were in error or why they were correct. Only in this way can we gain confidence in a forecasting procedure.

It is unlikely that one can foresee the future with a high degree of accuracy. A general rule for sound forecasting is, therefore: *revise any forecast continuously*. Forecasting is not a one-shot operation. To be effective, it requires continuous attention. Thus, if the demand for a company's product for 1965 is estimated in 1958, a good rule would be to revise the estimate in 1959, 1960, and so on. Unanticipated developments will often change our picture of the future, or at least clarify it. In terms of any original decisions and actions that have been taken, this rule implies continuous modification, where possible. Decisions about the future cannot be made and then not reviewed. Such decisions must be made sequentially and continuously.

In conclusion, remember that in projecting trends we are concerned with only one of the systematic elements that economists attribute to time series. Therefore, at times, the possible impact of the other elements—seasonal, cyclical, and random—must be considered.

The Moving Average. At times it is difficult to determine the nature of the trend because the trend movements are obscured by the effect of random and cyclical movements. A useful device for smoothing out the random fluctuations and the cyclical movements is a *moving average*. The method of computing a moving average is best explained with an illustration.

Example 14.10 Below a 3-item moving average is determined for the number of ordinary life insurance policies in force in the United States from 1950–57.

To take a 3-item moving average, we total the first 3 items ($64 + 67 + 70 = 201$). The total is placed opposite the middle item—201 is placed opposite 67. We then drop the first item (64) and add the fourth item (73) to get our second moving total ($67 + 70 + 73 = 210$). We then drop the second item (67) and add the fifth item (76) to get the third moving total ($70 + 73 + 76 = 219$), and so on. To get the 3-item moving average we divide each moving total by 3.

WORK SHEET FOR DETERMINING THREE-ITEM MOVING AVERAGE

Year	Number of Policies (in Millions)	3-Item Moving Total	3-Item Moving Average
1950.....	64		
1951.....	67	201	67.0
1952.....	70	210	70.0
1953.....	73	219	73.0
1954.....	76	229	76.3
1955.....	80	239	79.7
1956.....	83	250	83.3
1957.....	87		

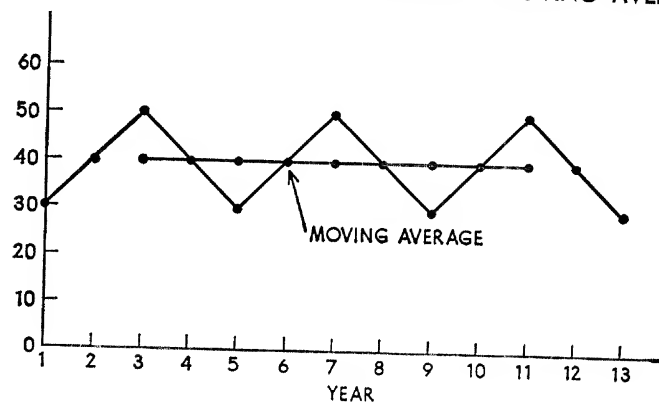
Observe that when we take a moving average, we lose observations at the beginning and end of a series. Thus, the original series had 8 items; the moving average series has only 6 items. This fact is a shortcoming of the use of the moving average for forecasting. We must not only project a series of moving averages into the future but we must first project the series into the present.

Now that we know how to determine a moving average, let us examine how a moving average can smooth fluctuations in a series of data.

Example 14.11 A hypothetical time series is plotted in Figure 14.7

Figure 14.7

HYPOTHETICAL TIME SERIES AND FOUR-YEAR MOVING AVERAGE



You will observe that the given data are periodic in nature. The number of years from peak to peak (year 3 to year 7 and year 7 to year 11) and from trough to trough (year 1 to year 5, year 5 to year 9, and year 9 to year 13) always is 4 years. Let us see what happens to the data if we take a moving average equal in length to the period of fluctuation—4 periods. The 4-item moving average for the data is computed below:

WORK SHEET FOR COMPUTING FOUR-ITEM MOVING AVERAGE

(1) Year	(2) Value	(3) 4-Item Moving Total	(4) 4-Item Moving Average	(5) 2-Item Moving Total	(6) 2-Item Moving Average
1.....	30				
2.....	40	160	40		
3.....	50	160	40	80	40
4.....	40	160	40	80	40
5.....	30	160	40	80	40
6.....	40	160	40	80	40
7.....	50	160	40	80	40
8.....	40	160	40	80	40
9.....	30	160	40	80	40
10.....	40	160	40	80	40
11.....	50	160	40	80	40
12.....	40				
13.....	30				

Note that for a 4-item moving average the middle of each group is between the second and third items of that group. Whenever an even-item moving average is taken, each moving average value falls between 2 items. It is desirable at times, therefore, to center the moving average. This can be done by taking a 2-item moving average of the moving average series already computed. The procedure is demonstrated in columns 5 and 6 above. The first value of the final moving average series is centered at period 3 and the last value is at period 11.

In the previous illustration, the 4-item moving average completely smooths out the ups and downs in the series which has a re-occurring cycle of 4 years. However, random movements and cycli-

cal movements are, as a rule, not completely periodic in nature. Therefore, most time series will not be smoothed perfectly by a moving average—some irregularities will remain. In general, an n -item moving average will have a smoothing effect on fluctuations with average periodicities of n time periods and less than n time periods. Thus, a 12-item moving average will smooth out fluctuations of 12 or fewer time periods in duration.

The moving average device is helpful in revealing the trend by smoothing out cyclical and random fluctuations. However, it is also possible for a moving average to introduce cyclical fluctuations into a series of data that are random in nature. This tendency is known as the Slutsky-Yule effect, named after the two men who studied this characteristic of moving averages. The possibility of this effect calls for great care on the part of the analyst who uses moving averages to smooth data.

Example 14.12 This illustration has been designed to show how a moving average can introduce the appearance of a cyclical fluctuation in time series. Let us assume that the following hypothetical data are given and that a 5-year moving average is computed.

Year	Value	5-Year Moving Total	5-Year Moving Average
1.....	100		
2.....	100		
3.....	100	500	100
4.....	100	420	84
5.....	100	420	84
6.....	20	420	84
7.....	100	420	84
8.....	100	420	84
9.....	100	500	100
10.....	100	500	100
11.....	100		
12.....	100		

Figure 14.8 portrays the original time series and the moving average.

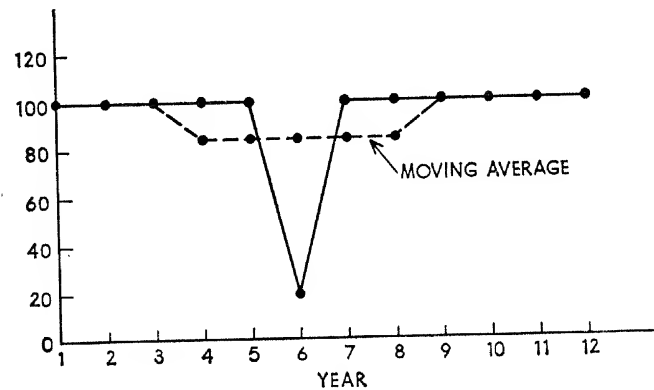
We thus observe a series which is constant except for the occurrence of a random event. However, the 5-year moving average has converted this into a series which appears to have a cycle with a flattened trough between years 4 and 8.

It is not unusual to find such a situation in a time series. Thus, the series might represent the volume of production of a given industry. Production is fairly constant for 5 years. During the sixth year a prolonged strike (random event) occurs and production drops off

sharply. The strike is settled, and then production reverts to its earlier level. A moving average, as we have seen, would make this random event appear as a cyclical fluctuation.

Figure 14.8

HYPOTHETICAL TIME SERIES AND A FIVE-YEAR MOVING AVERAGE



14.6 The Analysis of Seasonal Movements

Time series data fluctuate from month to month during any given year. For example, the demand for automobiles varies from month to month, as does the level of automobile production. When the pattern of variability remains fairly constant from year to year, we say that there is a systematic seasonal factor influencing the data. Thus, department store sales are high before Christmas and Easter and low during the summer months. The production of ladies' coats and suits is at a high level during the summer months and during February and March but is at a lower level during the rest of the year.

A knowledge of the seasonal pattern is an important factor in business planning. The knowledge that flower sales will be high in May because of Mother's Day tells the florist that he must stock flowers to meet the demand. The knowledge that production will be high during the summer months helps the garment manufacturer to plan his working capital requirements, recruitment of labor, machine maintenance needs, and so on.

Knowledge of seasonal movements also helps management to evaluate past and current performance. Thus, in a given year, sales for the month of March might be 10 per cent above February sales but less than the 20 per cent increase that should be expected on a

seasonal basis. An alert firm should investigate the reasons for such occurrences and, if they find it necessary, should make the appropriate policy changes.

A business, knowing the character of a seasonal pattern, has a basis for deciding, if possible, whether or not to change it. Thus, one soap manufacturing concern was able to smooth out the ups and downs in its pattern of production by changing its method of distribution. A cold cereal manufacturer, observing that cereal sales fall off during winter months, might attempt to raise the level of sales during that period by offering special prizes to purchasers during the winter months.

At other times, it is desirable to measure seasonal fluctuations simply to eliminate them from the original data. This often enables us to study the other systematic movements (trend and cycle) more easily.

Various methods have been developed to measure seasonal variation. The logic behind all of these methods is to eliminate the influence of trend, cyclical, and random elements from the original time series data so that only the influence of the seasonal element remains. We shall describe only the *ratio-to-moving-average method* for isolating seasonality, since it is most widely applicable.

The basic steps in the "ratio-to-moving-average" method to derive a monthly seasonal are as follows:

1. First, take a 12-month moving average of the data. As you will recall, a 12-month moving average smooths out fluctuations of 12 months or less duration—these will include seasonal fluctuations. Thus, our first step in isolating seasonal movements is to eliminate them from the data. Again assume that an observation at time t is the product of four elements,

$$Y_t = T_t \times C_t \times S_t \times R_t.$$

The purpose of the moving average is to eliminate the seasonal movements from the data. The moving average, therefore, is approximately equal to

$$T_t \times C_t \times R_t.$$

2. The second step is to take the ratio of the value of each month to the moving average for that month. The approximate effect of this step can be stated symbolically as follows:

$$\frac{Y_t}{\text{MOVING AVERAGE}} = \frac{T_t \times C_t \times S_t \times R_t}{T_t \times C_t \times R_t} = S_t.$$

Thus, this step gives an approximation of the seasonal movement in each month.

3. The ratio-to-moving-average terms for all Januaries, Februaries, . . . , Decembers are then averaged (and adjusted) to determine a set of 12 seasonal indices. The effect of the averaging step is to remove random, cyclical, or trend elements that are not removed in the earlier steps.

In the example that follows we apply this method to estimate a monthly seasonal index for the number of initial claims for unemployment insurance made on state programs. Unemployment data are one of the key indicators of shifts in general business activity. However, to reach decisions on the basis of unemployment data, it is necessary to adjust the data for seasonal fluctuations. Before further discussion, let us compute seasonal indices.

Example 14.13 Table 14.2 presents a worksheet for determining the seasonal movement in initial unemployment insurance claims made on state programs by the ratio-to-moving-average method. In this illustration, data for only 5 years, 1953 to 1957, are used to compute seasonal indices. This is, as a rule, too short a period since the ratio-to-moving-average method relies on averaging to remove nonseasonal effects. The longer the series, the greater the likelihood of averaging out nonseasonal effects. Generally, 8 to 15 years has been found to be a satisfactory period.

The first step in the ratio-to-moving-average method, the computation of the 12-month moving average, is performed in columns (b) and (c) of Table 14.2. You will notice that we have centered the moving average opposite the seventh month (July) instead of between the sixth and seventh months. In this way we avoid the necessity of taking a 2-item moving average to center our original moving averages on a month instead of between months. In principle, it may be preferable to follow the more detailed procedure. However, such refinements result in negligible improvements and usually are not worth the added cost. The moving average is presented with the original data in Figure 14.9 on page 476.

The second step, the computation of the ratios of the actual data to the moving average, is performed in column (d). This yields an approximation of the seasonal fluctuation for each month.

The third step, the averaging of the ratios for each month is performed in Table 14.3. In averaging the ratios we should try to avoid the distortion of the average by extreme values. This may be done by computing the median, which is unaffected by extreme values.

An alternate method is to take a *medial average*, generally the mean of the middle three or five ratios. A *medial average*, the mean

Table 14.2

WORK SHEET FOR DETERMINING MONTHLY RATIO TO MOVING
AVERAGE FOR INITIAL UNEMPLOYMENT INSURANCE CLAIMS
ON STATE PROGRAMS FOR THE YEARS 1953-57

Year and Month	(a) Weekly Average of Number of Initial Claims (in Thou- sands)	(b) 12-Month Moving Total	(c) 12-Month Moving Average	(d) Ratio to Moving Average $(a \div c) \times 100$
1953:				
January.....	236			
February.....	184			
March.....	179			
April.....	190			
May.....	186			
June.....	182			
July.....	213	2,601	217	98.2
August.....	189	2,781	232	81.5
September.....	186	2,932	244	76.2
October.....	209	3,056	251	83.2
November.....	296	3,194	266	111.3
December.....	351	3,300	275	127.6
1954:				
January.....	416	3,407	284	146.7
February.....	335	3,497	291	115.1
March.....	303	3,571	298	111.7
April.....	328	3,640	303	108.3
May.....	292	3,693	313	93.3
June.....	289	3,668	306	94.9
July.....	303	3,632	303	100.0
August.....	263	3,572	298	88.3
September.....	255	3,494	291	87.6
October.....	262	3,407	284	92.3
November.....	271	3,317	276	98.2
December.....	315	3,230	269	113.4
1955:				
January.....	356	3,143	262	135.9
February.....	257	3,068	256	100.0
March.....	216	2,994	250	86.4
April.....	238	2,902	242	98.3
May.....	205	2,827	236	87.7
June.....	202	2,767	231	87.4
July.....	228	2,722	227	100.0
August.....	189	2,673	223	84.8
September.....	163	2,666	222	73.4
October.....	187	2,663	222	84.2
November.....	211	2,659	222	95.0
December.....	270	2,670	223	121.1
1956:				
January.....	307	2,673	223	137.7
February.....	250	2,699	225	111.1

Table 14.2—Continued

Year and Month	(a) Weekly Average of Number of Initial Claims (in Thou- sands)	(b) 12-Month Moving Total	(c) 12-Month Moving Average	(d) Ratio to Moving Average $(a \div c) \times 100$
March.....	213	2,692	224	95.1
April.....	234	2,719	227	103.1
May.....	216	2,713	226	95.6
June.....	205	2,723	227	90.3
July.....	254	2,740	229	110.9
August.....	182	2,779	231	78.8
September.....	190	2,780	232	81.9
October.....	181	2,781	232	78.0
November.....	221	2,797	233	94.8
December.....	293	2,799	233	125.7
1957:				
January.....	340	2,814	235	144.7
February.....	251	2,836	236	106.4
March.....	214	2,845	237	90.3
April.....	250	2,901	242	103.3
May.....	218	2,979	248	87.9
June.....	220	3,078	257	85.6
July.....	276	3,230	269	102.6
August.....	191			
September.....	246			
October.....	259			
November.....	320			
December.....	445			

Source: *Economic Report of the President*, January, 1956, p. 187, and January, 1958, p. 139.

of the items left after discarding the highest and the lowest ratios, is taken of the ratios in Table 14.3.

A last step requires the adjustment of the averages of the ratios (crude seasonal) so that they average 100 per cent each—that is, so that they total 1,200. The crude seasonal indices total only 1,192.9. Thus, each of these is multiplied by $(1,200/1,192.9) = 1.0059$. The resulting values are the seasonal indices which sum to 1,200.0. Table 14.3 also indicates this step.

The seasonal index for January, as computed in the last example, is 137.6. What does this mean? The index tells us that on the basis of the 1953–57 data, the weekly number of unemployment insurance claims for January averaged 37.6 per cent above the average for the year. The index of 84.6 for September indicates that claims for that month averaged 15.4 per cent below the average for the year.

Table 14.3
COMPUTATION OF SEASONAL INDEX FROM MONTHLY RATIOS TO MOVING AVERAGE

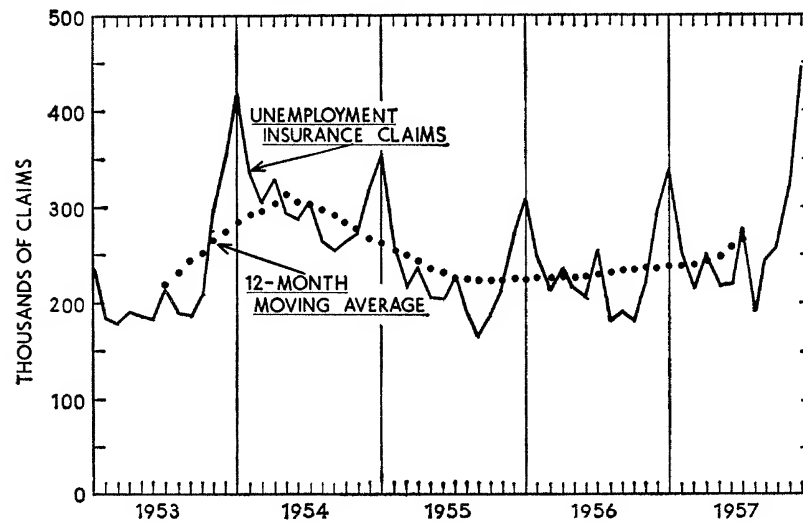
Year	Jan.	Feb.	March	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Total
1953.....	146.7	115.1	111.7	108.3	93.3	94.9	98.2	81.5	76.2	83.2	111.3	127.6	
1954.....	135.9	100.0	86.4	98.3	87.7	87.4	100.0	88.3	87.6	92.3	98.2	113.4	
1955.....	137.7	111.1	95.1	103.1	95.6	90.3	110.9	84.8	73.4	84.2	95.0	121.1	
1956.....	144.7	106.4	90.3	103.3	87.9	85.6	102.6	78.8	81.9	78.0	94.8	125.7	
Medial Average*.....	136.8	108.8	92.7	103.2	90.6	88.9	100.9	83.2	84.1	83.7	96.6	123.4	1,192.9
Seasonal Index†.....	137.6	109.5	93.3	103.8	91.1	89.4	101.5	83.7	84.6	84.2	97.2	124.1	1,200.0

* The average excluding the highest and lowest ratio.

† Adjustment factor = $\frac{1,200}{1,192.9} = 1.0059$.

Figure 14.9

INITIAL UNEMPLOYMENT INSURANCE CLAIMS ON STATE PROGRAMS,
1953-57
(Weekly Average in Thousands)



To adjust data for seasonal influence, actual data observations are divided by the seasonal index. Thus, symbolically

$$\begin{aligned}
 \text{SEASONALLY ADJUSTED DATA} &= \frac{Y_t}{S_t} \\
 &= \frac{T_t \times S_t \times C_t \times R_t}{S_t} \\
 &= T_t \times C_t \times R_t.
 \end{aligned}$$

Data which have been adjusted for seasonal variation do not show what the behavior of the series would have been in the absence of seasonal forces. The seasonal forces make up part of a complex of factors influencing the series. How the series would have behaved if all other forces but the seasonal were present is a question no one can answer. The deseasonalized data represent an average value. For example, if the remainder of the year would have operated on the same level as the month of January, then the average number of initial claims for the year 1957 would have been

$$\begin{aligned}
 \frac{\text{JANUARY, 1957, AVERAGE WEEKLY CLAIMS}}{\text{SEASONAL INDEX}} &= \frac{340}{137.6} \\
 &= 247 \text{ THOUSAND CLAIMS.}
 \end{aligned}$$

Example 14.14 During the recession of 1957-58 the President and his economic advisors studied unemployment figures carefully as an indicator of the state of the economy in order to reach decisions on anti-recession measures. In addition to the unemployment figures relating to workers covered by state unemployment insurance programs, there is the monthly estimate of unemployment based on a sample survey, conducted by the Department of Commerce. For February, 1958, this survey reported 5,173,000 unemployed. The March figure was 5,198,000 unemployed—indicating an actual increase of only 25,000. However, to properly assess the change, the two figures must be adjusted for seasonal variation. When this is done by dividing each figure by the appropriate seasonal index, the increase in unemployment on a seasonally adjusted basis is more than 200,000. The calculations, based on Department of Commerce seasonal indices for February and March, 1958, are as follows:

Month	Actual Number of Unemployed (in Thousands)	Seasonal Index	Seasonally Ad- justed Number of Unemployed (in Thousands)
February.....	5,173	115.9	4,463
March.....	5,198	111.3	4,670
Increase.....	25		207

The seasonal index considered in Example 14.13 is called a *constant seasonal*. That is, the seasonal index for each month is assumed to be the same from year to year. This very often is not a valid assumption since seasonal indices sometimes exhibit trends and cyclical movements. Such indices are called *moving seasonals*. The method we have described, with certain additional steps, can be applied to determine a moving seasonal, although we shall not cover the technique in this text.

The index we have computed also can be described as a *relative seasonal* since it is expressed as a percentage. A relative seasonal assumes that the seasonal effect is proportional to the level of the series. Thus, at higher levels the amplitude of seasonal fluctuations will be greater. The seasonal index for February, in Example 14.13, was 109.6. This implies that the actual values are, on the average, 9.6 per cent above $C_t \times T_t \times R_t$. At times, the seasonal fluctuations are not proportional to the level of the series but remain approximately the same in absolute amount no matter what the level. A seasonal index which describes such movements is called an *absolute seasonal*. The method for deriving this type of seasonal index differs only slightly from the method described above.

Seasonal indices, therefore, can be either constant or moving. In either event they may be stated in relative or absolute terms. As you can well imagine, the computation of a seasonal index is a laborious and costly process. However, during the past few years, electronic computers have been used to compute seasonal indices. The procedure the machine simulates is the ratio-to-moving-average method. It can be used to compute relative or absolute seasonals, and either constant or moving seasonals. The machine not only computes the seasonal index but also adjusts the original series.

14.7 Cyclical Fluctuations

The classical method for isolating cyclical fluctuations is called the *residual method*. The trend and seasonal indices are determined and then eliminated. This leaves the cyclical and random effects. A moving average, usually of from 3 to 5 items, then is used to smooth out the random influences. The resulting values are considered measures of cyclical fluctuations.

There are serious objections to this method. These were stated most effectively by Arthur F. Burns and Wesley C. Mitchell, two authorities in the empirical study of business cycles:¹

We do not follow that plan. In the first place, the isolation of cyclical fluctuations is a highly uncertain operation. Edwin Frickey once diligently assembled 23 trend lines fitted by various investigators to pig iron production in the United States, and found that some of the trend lines yield cycles averaging 3 or 4 years in duration while others yield cycles more than ten times as long. This range of results illustrates vividly the uncertainty that attaches to separations of trends and cycles, though it perhaps exaggerates the difficulties. If an investigator fits a trend line in a mechanical manner, without specifying in advance his conception of the secular trend or of cyclical fluctuations, he may get "cycles" of almost any duration. But an informed investigator who is seriously studying cycles of a given order of duration will use whatever guidance he can get from history and statistics; he will scrutinize the movements of the original data, seek to mark off in advance the cycles or traces of cycles that correspond to his basic conception, then choose a trend line that cuts through and exposes the cycles in which his interest centers. Yet this procedure also illustrates the difficulties of segregating trends and cycles. For it leaves room for choice of the trend line, the method of fit, and the method of trend elimination. Further, it makes the trend depend upon the cycles, and may not lead to the discovery of cycles that are obscured by the trend. To judge what features of the

¹ Arthur F. Burns and Wesley C. Mitchell, *Measuring Business Cycles* (New York: National Bureau of Economic Research, 1946), pp. 37-38. Quoted by permission.

final result are merely technical and what features are a significant characteristic of the series is likely to require considerable testing and experimenting even on the part of a skilled technician.

It is fairly common for statisticians to assume that the elimination of the secular trend from a time series indicates what the course of the series would have been in the absence of secular movements, and that the graduation of a time series, whether in original or trend-adjusted form, indicates what the course of the series would have been in the absence of random movements. There is no warrant for such simple interpretations. A "least squares" trend line fitted, for example, to grocery chain store sales in the United States may move majestically on a chart, but the analytic significance of the trend line is obscure. At least some of the "growth factors" impinging on this branch of business—the addition of meats and vegetables to the grocery line, the rise of supermarkets, special taxes on chain stores—have made their influence felt spasmodically. When a continuous "trend factor" is eliminated from the data, it is therefore difficult to say what influences impinging on the activity have been removed and what influences have been left in the series. Cyclical graduations are no easier to interpret than trend adjustments. Systematic smoothing of a time series will, indeed, eliminate short-run oscillations produced by random factors; but can it eliminate the influence of powerful random factors—such as a protracted strike, or a succession of bad harvests, or a great war?

There is always danger that the statistical operations performed on the original data may lead an investigator to bury real problems and worry about false ones. . . .

Mitchell, during his lifetime, was associated with the National Bureau of Economic Research. Burns still is associated with the NBER. This organization, since its creation in the middle 1920's, has been a guiding force in the conduct of empirical research relating to the business cycle. The Bureau has analyzed the cyclical behavior of over 1,000 economic time series, American and foreign. It prefers to use monthly or quarterly data to study cycles since annual data often conceal the timing of cyclical peaks or troughs. Data are adjusted for seasonal fluctuation before the study of cyclical movements begins. This facilitates the study of the cyclical movements which often are obscured by seasonal fluctuations.

The Bureau relates its study of cycles in each series to a table of reference dates, showing the months and years of troughs and peaks in general business activity. For the United States, the tables are based upon business annals and then refined, tested, and amended on the basis of subsequent research. The peaks and troughs in the cycles for each series are examined for leads and lags relative to the reference cycles. The Bureau at present is experimenting

with the information it has gathered relative to leads and lags to predict fluctuations in general business activity.

The National Bureau, in its empirical studies, has deviated from the traditional approach to the analysis of systematic movements in time series. Other investigators have set up methods of analysis that also differ from the one we have defined. A discussion of these alternative methods is beyond the scope of this book. We make reference to them only to stress that the traditional residual method has several alternatives.

While the various approaches to time series analysis have been useful in describing the historical behavior of time series, none of these has yielded reliable methods of predicting future cyclical behavior. In most projections on the basis of time series it is necessary to use judgment reinforced with various types of economic data to predict the nature of cyclical fluctuations. Thus, the best (and perhaps the only) way to forecast cyclical shifts in business activity is through a thorough understanding of existing economic conditions and an economic analysis of what they might imply for the immediate future.

14.8 Trend, Seasonal, and Cyclical Elements—a Synthesis

We have reviewed the methods used to isolate the systematic elements in a time series. How do we combine these elements for forecasting purposes?

In long-range forecasting we are concerned mainly with trend. If we need to predict demand for electric power 20 years hence, it is sufficient to predict the trend of annual demand. On the basis of this trend forecast a utility company, for example, could decide on the extent of an expansion program and make plans for its long-range financing. It would make little sense to attempt to predict the seasonal pattern of demand at that time since such data, even if they could be predicted, would be of little value in decision making today.

At times, it might be desirable to predict the cyclical influence in a long-range forecast. However, such predictions still are beyond our fondest hope. Even short-run cyclical effects cannot be predicted consistently. Therefore, as a practical expedient one should regularly check a long-range forecast of trend. As the period for which the forecast was made approaches, the revised trend forecast can be modified on the basis of a qualitative analysis of the cyclical influence.

In conclusion, let us now look at the following illustration of a short-run forecast.

Example 14.15 A company desires to forecast sales for the coming year. It can use the forecast as a basis for decisions on the budget for the year, production schedules, sales quotas, advertising plans, procurement policy, and labor needs. In the short-run forecast, however, it is necessary to know the seasonal pattern of sales during the year, in addition to forecasting annual sales. The Eastman Kodak Company prepares such short-run forecasts for its products.² Its procedure is as follows:

1. As a first step, the company extrapolates the past trend of sales. This gives an estimate of trend.
2. It is next necessary to modify the trend forecast for cyclical and possible random effects, and for possible changes in past trend forces in the future. The modifying factors considered are as follows:
 - a) The assumed pattern of general business conditions.
 - b) Any changes in the competitive position of the company.
 - c) The current dealer inventory situation.
 - d) Sales plans and policies.
 - e) Advertising and promotional plans.
 - f) Product research and development.
 - g) Price changes anticipated.
 - h) Capacity.
 - i) Availability of labor and materials.
 - j) Government controls.

The analysis of the influence of these factors on past trends is the basis for the adjustment.

It then is necessary to allocate the annual forecast among the months of the year on the basis of the seasonal pattern. Let us assume that the annual sales forecast for a product is \$1,800,000. The average monthly sales will, therefore, be \$150,000. If the seasonal index for January is .80, then the forecast of January sales is:

$$\$150,000 \times .80 = \$120,000.$$

Other monthly sales estimates are derived by multiplying \$150,000 by the appropriate seasonal indices.

This illustration again should emphasize that forecasting on the basis of time series analysis leans heavily on expert judgment. The projection of the trend is a relatively minor step in the annual sales forecast. The subsequent adjustment of this projection is the key factor in determining the actual forecast. Unfortunately, predictions of this type are not wholly within the province of statistical theory. The risk of an error in the forecasting procedure cannot be meas-

² See *Company Organization for Economic Forecasting*, American Management Association, Research Report No. 28 (New York, 1957), p. 51.

ured by probability theory, and, hence, such risks cannot be controlled.

14.9 You Should Now Know That

Forecasting involves the methods and theory of making decisions for the future on the basis of a study of present populations.

A time series consists of a set of observations made at different periods of time.

At the present state of our knowledge, any forecast of the future behavior of a time series must rest more heavily on judgment than on statistical theory.

A time series may be graphically presented on arithmetic scale charts or on semilogarithmic charts.

The trend of a time series is the resultant of the long-term economic forces.

Cycles are oscillatory movements, of more than one year in duration, which result from ups and downs in general business activity.

Seasonal fluctuations are variations which occur regularly during the period of a year and are the result of climatic conditions and customs.

Time series are also subject to random fluctuations.

Time series analysis deals with the isolation of systematic elements—i.e., trend, cyclical, and seasonal movements.

On the basis of the factors identified, an attempt is made to forecast the future course of events.

A time series may be tested for trend on the basis of observed signs of differences between successive terms.

The trend of a time series may be indicated by a moving average.

A trend line may be fitted to a time series as a freehand line or as a mathematical equation.

Seasonal fluctuations may be estimated by the ratio-to-moving-average method.

Seasonal indices may be constant or moving and also may be relative or absolute.

Cyclical fluctuations are best forecasted on the basis of sound economic knowledge of the current state of the economy.

Every forecast involves assumptions; a good forecaster recognizes these assumptions explicitly.

Effective forecasting requires that any projection should be revised continuously and, where possible, requires continuous modification of original decisions.

14.10 You Should Now Be Able to Solve

1. For each of the following problems indicate what systematic component(s)—trend, seasonal, cyclical—of the time series must be measured to provide a basis for action.

- a) An investor wants to decide, on the basis of past price and earnings movements, whether or not to buy XYZ stock as a long-term commitment.
- b) A speculator wants to decide, on the basis of past price and earnings movements, whether or not to buy ABC stock for the possibility of a quick profit.
- c) An estimate of the demand for steel in 1975 is required as a basis for the decision on how much of an increase in capacity should be planned.
- d) A company must prepare its monthly production quotas for the coming year.
- e) A state wants to estimate the yield of a gasoline tax in the coming fiscal year for budgetary purposes.

2. Explain and discuss the following statement: Forecasting the future course of economic events is not wholly a statistical problem.

3. The following time series data are the average price per common share for Minnesota Mining and Manufacturing Company from 1946 to 1958:

Year	Average Price* (in Dollars)	Year	Average Price* (in Dollars)
1946.....	6.0	1953.....	25.6
1947.....	7.4	1954.....	36.3
1948.....	8.3	1955.....	48.8
1949.....	10.4	1956.....	64.1
1950.....	15.3	1957.....	79.5
1951.....	22.8	1958.....	94.8
1952.....	21.8		

* Average of high and low prices during calendar year; adjusted for stock splits through 1958.

Plot this time series on arithmetic scale graph paper and on semi-logarithmic scale graph paper. If you were to fit a freehand trend line to the data as a basis for prediction, which chart would facilitate your task more? Explain your answer.

What type of information would you need before you drew and projected a trend line?

4. The following time series represents a firm's ratios of operating income to net sales for 1937-58. Formulate a statistical decision rule in terms of a statistic measuring the number of positive dif-

ferences between successive observations to provide a basis for the decision on whether or not to fit a trend to the series. Apply the rule to the data. What is the decision?

Year	Ratio (in Per Cent)	Year	Ratio (in Per Cent)
1937.....	10.1	1948.....	10.7
1938.....	7.9	1949.....	7.6
1939.....	11.3	1950.....	10.6
1940.....	10.8	1951.....	9.8
1941.....	9.2	1952.....	8.9
1942.....	5.3	1953.....	9.3
1943.....	5.9	1954.....	10.3
1944.....	7.0	1955.....	9.9
1945.....	6.8	1956.....	9.9
1946.....	2.2	1957.....	8.7
1947.....	11.1	1958.....	6.8

5. The following time series lists the net sales of a large food company from 1947 through 1957 inclusive:

Year	Net Sales (Millions of Dollars)	Year	Net Sales (Millions of Dollars)
1947.....	112.9	1953.....	151.0
1948.....	119.8	1954.....	146.8
1949.....	90.0	1955.....	162.4
1950.....	98.6	1956.....	166.8
1951.....	125.1	1957.....	166.1
1952.....	117.0		

- Plot this time series on arithmetic scale graph paper.
- Fit a least-squares straight-line trend line to the data and draw the trend line on the chart constructed in part (a).
- Discuss the adequacy of the line as a basis for predicting sales in 1958; as a basis for predicting sales in 1968. Consider alternatives to simply extending the line.
- Based upon your discussion in part (c), prepare estimates of sales in 1958 and in 1968 for use by the controller in deciding on new financing. List all pertinent assumptions made in preparing these forecasts.
- Critically and constructively analyze the use of time series analysis in this type of problem.

6. To consider shifts in the level of sales due to factors other than the seasonal element, a sales executive might take the ratio of sales in January, 1959, to those in January, 1958, the ratio of sales in

February, 1959, to those in February, 1958, and so forth. Will these ratios be influenced by seasonal factors in any way? Explain your answer.

7. The following data represent the sales of a product by a plastics manufacturer, by months, from 1952 through 1958 inclusive:

MONTHLY SALES
(Thousands of Dollars)

	1952	1953	1954	1955	1956	1957	1958
January.....	14.3	15.5	15.2	16.1	17.4	17.9	18.3
February.....	13.9	15.2	15.2	16.0	17.0	17.1	17.8
March.....	14.7	15.5	15.8	16.9	17.8	18.4	19.2
April.....	13.7	15.3	15.1	16.1	17.1	17.4	17.9
May.....	14.0	14.2	14.8	15.6	17.0	16.2	18.0
June.....	13.2	14.3	13.6	15.2	16.8	16.4	17.5
July.....	12.0	12.1	14.1	14.8	15.8	16.3	16.9
August.....	13.3	14.1	13.4	15.0	16.4	16.8	18.5
September.....	14.4	15.2	16.0	16.9	17.7	18.4	19.2
October.....	15.3	16.2	16.8	17.7	18.9	20.1	20.8
November.....	14.3	15.1	16.2	17.0	17.9	18.2	20.1
December.....	14.2	15.0	15.8	16.5	17.3	18.0	18.8

- Plot this time series on arithmetic scale graph paper. Qualitatively describe the seasonal pattern.
- Compute monthly seasonal indices for these data. Do the results bear out your description of the seasonal pattern?
- Deseasonalize the data and plot the deseasonalized data on the chart prepared in part (a).
- If you were required to find the trend of this series, would you fit the trend line to the actual data or the deseasonalized data? Explain your answer.
- What assumptions would you make if you were to apply these indices after 1958?
- The company's sales quota for the product in 1959 is \$225,000. Estimate the monthly sales quotas on the basis of the seasonal indices you have computed.

14.11 You Will Also Find That

The following references are among those which consider some of the techniques that may be useful in time series analysis:

- CROXTON, FREDERICK E., and COWDEN, DUDLEY J. *Applied General Statistics*, pp. 240-392. 2d ed. New York: Prentice-Hall, Inc., 1955.
- MANDELL, B. J. *Statistics for Management*, pp. 277-301. Baltimore, Maryland: Dangary Publishing Co., 1956.
- MILLS, FREDERICK C. *Statistical Methods*, pp. 319-425. 3d ed. New York: Henry Holt & Co., Inc., 1955.

SPROWLS, R. CLAY. *Elementary Statistics*, pp. 262-331. New York: McGraw-Hill Book Co., Inc., 1955.

Some of the basic economic factors which must be considered in any forecasting problem are covered in the following texts:

BASSIE, V. LEWIS. *Economic Forecasting*. New York: McGraw-Hill Book Co., Inc., 1958.

BRATT, ELMER CLARK. *Business Cycles and Forecasting*. 4th ed. Homewood, Ill.: Richard D. Irwin & Co., Inc., 1953.

Appendixes

APPENDIX · A

Table of Squares, Square Roots, and Reciprocals $N = 1$ to $1,000^*$

The table given on the following ten pages will be found useful for numerical calculations. It includes squares, square roots, and reciprocals for all the integers from 1 through 1,000.

The table is particularly useful to determine square roots. For any given integer, N , the table gives value of

$$\sqrt{N} \quad \text{and} \quad \sqrt{10N}$$

Thus, for $N = 184$ the table indicates that

$$\sqrt{N} = \sqrt{184} = 13.56466$$

and that

$$\sqrt{10N} = \sqrt{1840} = 42.89522$$

The square roots of other numbers besides those of the integers given may be easily determined if relationships such as the following are kept in mind:

$$\sqrt{100N} = 10\sqrt{N}$$

$$\sqrt{1000N} = 10\sqrt{10N}$$

$$\sqrt{\frac{1}{10}N} = \frac{1}{10}\sqrt{10N}$$

$$\sqrt{\frac{1}{100}N} = \frac{1}{10}\sqrt{N}$$

* From Charles D. Hodgman (ed.), *C.R.C. Standard Mathematical Tables* (11th ed.; Cleveland: Chemical Rubber Publishing Company, 1957), by permission.

For example, note that

$$\begin{aligned}\sqrt{.0184} &= \sqrt{\frac{1}{10000}} (184) \\ &= \frac{1}{100} \sqrt{184}\end{aligned}$$

The table indicates that $\sqrt{184} = 13.56466$ and hence

$$\begin{aligned}\sqrt{.0184} &= \frac{1}{100} (13.56466) \\ &= .1356466\end{aligned}$$

Observe, if you use the table of reciprocals, that zeros immediately after the decimal point of the reciprocal usually are written, not in the body of the table, but rather in the column caption. For example, verify that the reciprocal of 184 is .005434783 (since the caption of the column $1/N$ lists .00 and that should precede all values in the column).

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 1-100

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$	N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.0
1	1	1.000 000	3.162 278	1.0000000	50	2 500	7.071 068	22.36068	2000000
2	4	1.414 214	4.472 136	.5000000	51	2 601	7.141 428	22.58318	1960784
3	9	1.732 051	5.477 226	.3333333	52	2 704	7.211 103	22.80351	1923077
4	16	2.000 000	6.324 555	.2500000	53	2 809	7.280 110	23.02173	1886792
5	25	2.236 068	7.071 068	.2000000	54	2 916	7.348 469	23.23790	1851852
6	36	2.449 490	7.745 967	.1666667	55	3 025	7.416 198	23.45208	1818182
7	49	2.645 751	8.366 600	.1428571	56	3 136	7.483 315	23.66432	1785714
8	64	2.828 427	8.944 272	.1250000	57	3 249	7.549 834	23.87467	1754386
9	81	3.000 000	9.486 833	.1111111	58	3 364	7.615 773	24.08319	1724138
10	100	3.162 278	10.00000	.1000000	59	3 481	7.681 146	24.28992	1694915
11	121	3.316 625	10.48809	.09090909	60	3 600	7.745 967	24.49490	1666667
12	144	3.464 102	10.95445	.08333333	61	3 721	7.810 250	24.69818	1639344
13	169	3.605 551	11.40175	.07692308	62	3 844	7.874 008	24.89980	1612903
14	196	3.741 657	11.83216	.07142857	63	3 969	7.937 254	25.09980	1587302
15	225	3.872 983	12.24745	.06666667	64	4 096	8.000 000	25.29822	1562500
16	256	4.000 000	12.64911	.06250000	65	4 225	8.062 258	25.49510	1538462
17	289	4.123 106	13.03840	.05882353	66	4 356	8.124 038	25.69047	1515152
18	324	4.242 641	13.41641	.05555556	67	4 489	8.185 353	25.88436	1492537
19	361	4.358 899	13.78405	.05263158	68	4 624	8.246 211	26.07681	1470588
20	400	4.472 136	14.14214	.05000000	69	4 761	8.306 624	26.26785	1449276
21	441	4.582 576	14.49138	.04761905	70	4 900	8.366 600	26.45751	1428571
22	484	4.690 416	14.83240	.04545455	71	5 041	8.426 150	26.64583	1408451
23	529	4.795 832	15.16575	.04347826	72	5 184	8.485 281	26.83282	1388889
24	576	4.898 979	15.49193	.04166667	73	5 329	8.544 004	27.01851	1369863
25	625	5.000 000	15.81139	.04000000	74	5 476	8.602 325	27.20294	1351351
26	676	5.099 020	16.12452	.03846154	75	5 625	8.660 254	27.38613	1333333
27	729	5.196 152	16.43168	.03703704	76	5 776	8.717 798	27.56810	1315789
28	784	5.291 503	16.73320	.03571429	77	5 929	8.774 964	27.74887	1298701
29	841	5.385 165	17.02939	.03448276	78	6 084	8.831 761	27.92848	1282051
30	900	5.477 226	17.32051	.03333333	79	6 241	8.888 194	28.10694	1265823
31	961	5.567 764	17.60682	.03225806	80	6 400	8.944 272	28.28427	1250000
32	1 024	5.656 854	17.88854	.03125000	81	6 561	9.000 000	28.46050	1234568
33	1 089	5.744 563	18.16590	.03030303	82	6 724	9.055 385	28.63564	1219512
34	1 156	5.830 952	18.43909	.02941176	83	6 889	9.110 434	28.80972	1204819
35	1 225	5.916 080	18.70829	.02857143	84	7 056	9.165 151	28.98275	1190476
36	1 296	6.000 000	18.97367	.02777778	85	7 225	9.219 544	29.15476	1176471
37	1 369	6.082 763	19.23538	.02702703	86	7 396	9.273 618	29.32576	1162791
38	1 444	6.164 414	19.49359	.02631579	87	7 569	9.327 379	29.49576	1149425
39	1 521	6.244 998	19.74842	.02564103	88	7 744	9.380 832	29.66479	1136364
40	1 600	6.324 555	20.00000	.02500000	89	7 921	9.433 981	29.83287	1123596
41	1 681	6.403 124	20.24846	.02439024	90	8 100	9.486 833	30.00000	1111111
42	1 764	6.480 741	20.49390	.02380952	91	8 281	9.539 392	30.16621	1098901
43	1 849	6.557 439	20.73644	.02325581	92	8 464	9.591 663	30.33150	1086957
44	1 936	6.633 250	20.97618	.02272727	93	8 649	9.643 651	30.49590	1075269
45	2 025	6.708 204	21.21320	.02222222	94	8 836	9.695 360	30.65942	1063830
46	2 116	6.782 330	21.44761	.02173913	95	9 025	9.746 794	30.82207	1052632
47	2 209	6.855 655	21.67948	.02127660	96	9 216	9.797 959	30.98387	1041667
48	2 304	6.928 203	21.90890	.02083333	97	9 409	9.848 858	31.14482	1030928
49	2 401	7.000 000	22.13594	.02040816	98	9 604	9.899 495	31.30495	1020408
50	2 500	7.071 068	22.36068	.02000000	99	9 801	9.949 874	31.46427	1010101
					100	10 000	10.00000	31.62278	1000000

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 100-200

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.0	N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00
100	10 000	10.00000	31.62278	10000000	150	22 500	12.24745	38.72983	6666667
101	10 201	10.04988	31.78050	09900990	151	22 801	12.28821	38.85872	6622517
102	10 404	10.09950	31.93744	09803922	152	23 104	12.32883	38.98718	6578947
103	10 609	10.14889	32.09361	09708738	153	23 409	12.36932	39.11521	6535948
104	10 816	10.19804	32.24903	09615385	154	23 716	12.40967	39.24283	6493506
105	11 025	10.24695	32.40370	09523810	155	24 025	12.44990	39.37004	6451613
106	11 236	10.29563	32.55764	09433962	156	24 336	12.49000	39.49684	6410256
107	11 449	10.34408	32.71085	09345794	157	24 649	12.52996	39.62323	6369427
108	11 664	10.39230	32.86335	09259259	158	24 964	12.56981	39.74921	6329114
109	11 881	10.44031	33.01515	09174312	159	25 281	12.60952	39.87480	6289308
110	12 100	10.48809	33.16625	09090909	160	25 600	12.64911	40.00000	6250000
111	12 321	10.53565	33.31666	09009009	161	25 921	12.68858	40.12481	6211180
112	12 544	10.58301	33.46640	08928571	162	26 244	12.72792	40.24922	6172840
113	12 769	10.63015	33.61547	08849558	163	26 569	12.76715	40.37326	6134969
114	12 996	10.67708	33.76389	08771930	164	26 896	12.80625	40.49691	6097561
115	13 225	10.72381	33.91165	08695652	165	27 225	12.84523	40.62019	6060606
116	13 456	10.77033	34.05877	08620690	166	27 556	12.88410	40.74310	6024096
117	13 689	10.81665	34.20526	08547009	167	27 889	12.92285	40.86563	5988024
118	13 924	10.86278	34.35113	08474576	168	28 224	12.96148	40.98780	5952381
119	14 161	10.90871	34.49638	08403361	169	28 561	13.00000	41.10961	5917160
120	14 400	10.95445	34.64102	08333333	170	28 900	13.03840	41.23106	5882353
121	14 641	11.00000	34.78505	08264463	171	29 241	13.07670	41.35215	5847953
122	14 884	11.04536	34.92850	08196721	172	29 584	13.11488	41.47288	5813953
123	15 129	11.09054	35.07136	08130081	173	29 929	13.15295	41.59327	5780347
124	15 376	11.13553	35.21363	08064516	174	30 276	13.19091	41.71331	5747126
125	15 625	11.18034	35.35534	08000000	175	30 625	13.22876	41.83300	5714286
126	15 876	11.22497	35.49648	07936508	176	30 976	13.26650	41.95235	5681818
127	16 129	11.26943	35.63706	07874016	177	31 329	13.30413	42.07137	5649718
128	16 384	11.31371	35.77709	07812500	178	31 684	13.34166	42.19005	5617978
129	16 641	11.35782	35.91657	07751938	179	32 041	13.37909	42.30839	5586592
130	16 900	11.40175	36.05551	07692308	180	32 400	13.41641	42.42641	5555556
131	17 161	11.44552	36.19392	07633588	181	32 761	13.45362	42.54409	5524862
132	17 424	11.48913	36.33180	07575758	182	33 124	13.49074	42.66146	5494505
133	17 689	11.53256	36.46917	07518797	183	33 489	13.52775	42.77850	5464481
134	17 956	11.57584	36.60601	07462687	184	33 856	13.56466	42.89522	5434783
135	18 225	11.61895	36.74235	07407407	185	34 225	13.60147	43.01163	5405405
136	18 496	11.66190	36.87818	07352941	186	34 596	13.63818	43.12772	5376344
137	18 769	11.70470	37.01351	07299270	187	34 969	13.67479	43.24350	5347594
138	19 044	11.74734	37.14835	07246377	188	35 344	13.71131	43.35897	5319149
139	19 321	11.78983	37.28270	07194245	189	35 721	13.74773	43.47413	5291005
140	19 600	11.83216	37.41657	07142857	190	36 100	13.78405	43.58899	5263158
141	19 881	11.87434	37.54997	07092199	191	36 481	13.82027	43.70355	5235602
142	20 164	11.91638	37.68289	07042254	192	36 864	13.85641	43.81780	5208333
143	20 449	11.95826	37.81534	06993007	193	37 249	13.89244	43.93177	5181347
144	20 736	12.00000	37.94733	06944444	194	37 636	13.92839	44.04543	5154639
145	21 025	12.04159	38.07887	06896552	195	38 025	13.96424	44.15880	5128205
146	21 316	12.08305	38.20995	06849315	196	38 416	14.00000	44.27189	5102041
147	21 609	12.12436	38.34058	06802721	197	38 809	14.03567	44.38468	5076142
148	21 904	12.16553	38.47077	06756757	198	39 204	14.07125	44.49719	5050505
149	22 201	12.20656	38.60052	06711409	199	39 601	14.10674	44.60942	5025126
150	22 500	12.24745	38.72983	06666667	200	40 000	14.14214	44.72136	5000000

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 200-300

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00	N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00
200	40 000	14.14214	44.72136	5000000	250	62 500	15.81139	50.00000	4000000
201	40 401	14.17745	44.83302	4975124	251	63 001	15.84298	50.09990	3984064
202	40 804	14.21267	44.94441	4950495	252	63 504	15.87451	50.19960	3968254
203	41 209	14.24781	45.05552	4926108	253	64 009	15.90597	50.29911	3952569
204	41 616	14.28286	45.16636	4901961	254	64 516	15.93738	50.39841	3937008
205	42 025	14.31782	45.27693	4878049	255	65 025	15.96872	50.49752	3921569
206	42 436	14.35270	45.38722	4854369	256	65 536	16.00000	50.59644	3906250
207	42 849	14.38749	45.49725	4830918	257	66 049	16.03122	50.69517	3891051
208	43 264	14.42221	45.60702	4807692	258	66 564	16.06238	50.79370	3875969
209	43 681	14.45683	45.71652	4784689	259	67 081	16.09348	50.89204	3861004
210	44 100	14.49138	45.82576	4761905	260	67 600	16.12452	50.99020	3846154
211	44 521	14.52584	45.93474	4739336	261	68 121	16.15549	51.08818	3831418
212	44 944	14.56022	46.04346	4716981	262	68 644	16.18641	51.18594	3816794
213	45 369	14.59452	46.15192	4694836	263	69 169	16.21727	51.28353	3802281
214	45 796	14.62874	46.26013	4672897	264	69 696	16.24808	51.38093	3787879
215	46 225	14.66288	46.36809	4651163	265	70 225	16.27882	51.47815	3773585
216	46 656	14.69694	46.47580	4629630	266	70 756	16.30951	51.57519	3759398
217	47 089	14.73092	46.58326	4608295	267	71 289	16.34013	51.67204	3745318
218	47 524	14.76482	46.69047	4587156	268	71 824	16.37071	51.76872	3731343
219	47 961	14.79865	46.79744	4566210	269	72 361	16.40122	51.86521	3717472
220	48 400	14.83240	46.90416	4545455	270	72 900	16.43168	51.96152	3703704
221	48 841	14.86607	47.01064	4524887	271	73 441	16.46208	52.05766	3690037
222	49 284	14.89966	47.11688	4504505	272	73 984	16.49242	52.15362	3676471
223	49 729	14.93318	47.22288	4484305	273	74 529	16.52271	52.24940	3663004
224	50 176	14.96663	47.32864	4464286	274	75 076	16.55295	52.34501	3649635
225	50 625	15.00000	47.43416	4444444	275	75 625	16.58312	52.44044	3636364
226	51 076	15.03330	47.53946	4424779	276	76 176	16.61325	52.53570	3623188
227	51 529	15.06652	47.64452	4405286	277	76 729	16.64332	52.63079	3610108
228	51 984	15.09967	47.74935	4385965	278	77 284	16.67333	52.72571	3597122
229	52 441	15.13275	47.85394	4366812	279	77 841	16.70329	52.82045	3584229
230	52 900	15.16575	47.95832	4347826	280	78 400	16.73320	52.91503	3571429
231	53 361	15.19868	48.06246	4329004	281	78 961	16.76305	53.00943	3558719
232	53 824	15.23155	48.16638	4310345	282	79 524	16.79286	53.10367	3546099
233	54 289	15.26434	48.27007	4291845	283	80 089	16.82260	53.19774	3533569
234	54 756	15.29706	48.37355	4273504	284	80 656	16.85230	53.29165	3521127
235	55 225	15.32971	48.47680	4255319	285	81 225	16.88194	53.38539	3508772
236	55 696	15.36229	48.57983	4237288	286	81 796	16.91153	53.47897	3496503
237	56 169	15.39480	48.68265	4219409	287	82 369	16.94107	53.57238	3484321
238	56 644	15.42725	48.78524	4201681	288	82 944	16.97056	53.66563	3472222
239	57 121	15.45962	48.88763	4184100	289	83 521	17.00000	53.75872	3460208
240	57 600	15.49193	48.98979	4166667	290	84 100	17.02939	53.85165	3448276
241	58 081	15.52417	49.09175	4149378	291	84 681	17.05872	53.94442	3436426
242	58 564	15.55635	49.19350	4132231	292	85 264	17.08801	54.03702	3424658
243	59 049	15.58846	49.29503	4115226	293	85 849	17.11724	54.12947	3412969
244	59 536	15.62050	49.39636	4098361	294	86 436	17.14643	54.22177	3401361
245	60 025	15.65248	49.49747	4081633	295	87 025	17.17556	54.31390	3389831
246	60 516	15.68439	49.59839	4065041	296	87 616	17.20465	54.40588	3378378
247	61 009	15.71623	49.69909	4048583	297	88 209	17.23369	54.49771	3367003
248	61 504	15.74802	49.79960	4032258	298	88 804	17.26268	54.58938	3355705
249	62 001	15.77973	49.89990	4016064	299	89 401	17.29162	54.68089	3344482
250	62 500	15.81139	50.00000	4000000	300	90 000	17.32051	54.77226	3333333

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 300-400

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00	N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00
300	90 000	17.32051	54.77226	3333333	350	122 500	18.70829	59.16080	2857143
301	90 601	17.34935	54.86347	3322259	351	123 201	18.73499	59.24525	2849003
302	91 204	17.37815	54.95453	3311258	352	123 904	18.76166	59.32959	2840909
303	91 809	17.40690	55.04544	3300330	353	124 609	18.78829	59.41380	2832861
304	92 416	17.43560	55.13620	3289474	354	125 316	18.81489	59.49790	2824859
305	93 025	17.46425	55.22681	3278689	355	126 025	18.84144	59.58188	2816901
306	93 636	17.49286	55.31727	3267974	356	126 736	18.86796	59.66574	2808989
307	94 249	17.52142	55.40758	3257329	357	127 449	18.89444	59.74948	2801120
308	94 864	17.54993	55.49775	3246753	358	128 164	18.92089	59.83310	2793296
309	95 481	17.57840	55.58777	3236246	359	128 881	18.94730	59.91661	2785515
310	96 100	17.60682	55.67764	3225806	360	129 600	18.97367	60.00000	2777778
311	96 721	17.63519	55.76737	3215434	361	130 321	19.00000	60.08328	2770083
312	97 344	17.66352	55.85696	3205128	362	131 044	19.02630	60.16644	2762431
313	97 969	17.69181	55.94640	3194888	363	131 769	19.05256	60.24948	2754821
314	98 596	17.72005	56.03570	3184713	364	132 496	19.07878	60.33241	2747253
315	99 225	17.74824	56.12486	3174603	365	133 225	19.10497	60.41523	2739726
316	99 856	17.77639	56.21388	3164557	366	133 956	19.13113	60.49793	2732240
317	100 489	17.80449	56.30275	3154574	367	134 689	19.15724	60.58052	2724796
318	101 124	17.83255	56.39149	3144654	368	135 424	19.18333	60.66300	2717391
319	101 761	17.86057	56.48008	3134796	369	136 161	19.20937	60.74537	2710027
320	102 400	17.88854	56.56854	3125000	370	136 900	19.23538	60.82763	2702703
321	103 041	17.91647	56.65686	3115265	371	137 641	19.26136	60.90977	2695418
322	103 684	17.94436	56.74504	3105590	372	138 384	19.28730	60.99180	2688172
323	104 329	17.97220	56.83309	3095975	373	139 129	19.31321	61.07373	2680965
324	104 976	18.00000	56.92100	3086420	374	139 876	19.33908	61.15554	2673797
325	105 625	18.02776	57.00877	3076923	375	140 625	19.36492	61.23724	2666667
326	106 276	18.05547	57.09641	3067485	376	141 376	19.39072	61.31884	2659574
327	106 929	18.08314	57.18391	3058104	377	142 129	19.41649	61.40033	2652520
328	107 584	18.11077	57.27128	3048780	378	142 884	19.44222	61.48170	2645503
329	108 241	18.13836	57.35852	3039514	379	143 641	19.46792	61.56298	2638522
330	108 900	18.16590	57.44563	3030303	380	144 400	19.49359	61.64414	2631579
331	109 561	18.19341	57.53260	3021148	381	145 161	19.51922	61.72520	2624672
332	110 224	18.22087	57.61944	3012048	382	145 924	19.54483	61.80615	2617801
333	110 889	18.24829	57.70615	3003003	383	146 689	19.57039	61.88699	2610966
334	111 556	18.27567	57.79273	2994012	384	147 456	19.59592	61.96773	2604167
335	112 225	18.30301	57.87918	2985075	385	148 225	19.62142	62.04837	2597403
336	112 896	18.33030	57.96551	2976190	386	148 996	19.64688	62.12890	2590674
337	113 569	18.35756	58.05170	2967359	387	149 769	19.67232	62.20932	2583979
338	114 244	18.38478	58.13777	2958580	388	150 544	19.69772	62.28965	2577320
339	114 921	18.41195	58.22371	2949853	389	151 321	19.72308	62.36986	2570694
340	115 600	18.43909	58.30952	2941176	390	152 100	19.74842	62.44998	2564103
341	116 281	18.46619	58.39521	2932551	391	152 881	19.77372	62.52999	2557545
342	116 964	18.49324	58.48077	2923977	392	153 664	19.79899	62.60990	2551020
343	117 649	18.52026	58.56620	2915452	393	154 449	19.82423	62.68971	2544529
344	118 336	18.54724	58.65151	2906977	394	155 236	19.84943	62.76942	2538071
345	119 025	18.57418	58.73670	2898551	395	156 025	19.87461	62.84903	2531646
346	119 716	18.60108	58.82176	2890173	396	156 816	19.89975	62.92853	2525253
347	120 409	18.62794	58.90671	2881844	397	157 609	19.92486	63.00794	2518892
348	121 104	18.65476	58.99152	2873563	398	158 404	19.94994	63.08724	2512563
349	121 801	18.68154	59.07622	2865330	399	159 201	19.97498	63.16645	2506266
350	122 500	18.70829	59.16080	2857143	400	160 000	20.00000	63.24555	2500000

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 400-500

<i>N</i>	<i>N</i> ²	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00	<i>N</i>	<i>N</i> ²	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00
400	160 000	20.00000	63.24555	2500000	450	202 500	21.21320	67.08204	2222222
401	160 801	20.02498	63.32456	2493766	451	203 401	21.23676	67.15653	2217295
402	161 604	20.04994	63.40347	2487562	452	204 304	21.26029	67.23095	2212389
403	162 409	20.07486	63.48228	2481390	453	205 209	21.28380	67.30527	2207506
404	163 216	20.09975	63.56099	2475248	454	206 116	21.30728	67.37952	2202643
405	164 025	20.12461	63.63961	2469136	455	207 025	21.33073	67.45369	2197802
406	164 836	20.14944	63.71813	2463054	456	207 936	21.35416	67.52777	2192982
407	165 649	20.17424	63.79655	2457002	457	208 849	21.37756	67.60178	2188184
408	166 464	20.19901	63.87488	2450980	458	209 764	21.40093	67.67570	2183406
409	167 281	20.22375	63.95311	2444988	459	210 681	21.42429	67.74954	2178649
410	168 100	20.24846	64.03124	2439024	460	211 600	21.44761	67.82330	2173913
411	168 921	20.27313	64.10928	2433090	461	212 521	21.47091	67.89698	2169197
412	169 744	20.29778	64.18723	2427184	462	213 444	21.49419	67.97058	2164502
413	170 569	20.32240	64.26508	2421308	463	214 369	21.51743	68.04410	2159827
414	171 396	20.34699	64.34283	2415459	464	215 296	21.54066	68.11755	2155172
415	172 225	20.37155	64.42049	2409639	465	216 225	21.56386	68.19091	2150538
416	173 056	20.39608	64.49806	2403846	466	217 156	21.58703	68.26419	2145923
417	173 889	20.42058	64.57554	2398082	467	218 089	21.61018	68.33740	2141328
418	174 724	20.44505	64.65292	2392344	468	219 024	21.63331	68.41053	2136752
419	175 561	20.46949	64.73021	2386635	469	219 961	21.65641	68.48357	2132196
420	176 400	20.49390	64.80741	2380952	470	220 900	21.67948	68.55655	2127660
421	177 241	20.51828	64.88451	2375297	471	221 841	21.70253	68.62944	2123142
422	178 084	20.54264	64.96153	2369668	472	222 784	21.72556	68.70226	2118644
423	178 929	20.56696	65.03845	2364066	473	223 729	21.74856	68.77500	2114165
424	179 776	20.59126	65.11528	2358491	474	224 676	21.77154	68.84766	2109705
425	180 625	20.61553	65.19202	2352941	475	225 625	21.79449	68.92024	2105263
426	181 476	20.63977	65.26868	2347418	476	226 576	21.81742	68.99275	2100840
427	182 329	20.66398	65.34524	2341920	477	227 529	21.84033	69.06519	2096436
428	183 184	20.68816	65.42171	2336449	478	228 484	21.86321	69.13754	2092050
429	184 041	20.71232	65.49809	2331002	479	229 441	21.88607	69.20983	2087683
430	184 900	20.73644	65.57439	2325581	480	230 400	21.90890	69.28203	2083333
431	185 761	20.76054	65.65059	2320186	481	231 361	21.93171	69.35416	2079002
432	186 624	20.78461	65.72671	2314815	482	232 324	21.95450	69.42622	2074689
433	187 489	20.80865	65.80274	2309469	483	233 289	21.97726	69.49820	2070393
434	188 356	20.83267	65.87868	2304147	484	234 256	22.00000	69.57011	2066116
435	189 225	20.85665	65.95453	2298851	485	235 225	22.02272	69.64194	2061856
436	190 096	20.88061	66.03030	2293578	486	236 196	22.04541	69.71370	2057613
437	190 969	20.90454	66.10598	2288330	487	237 169	22.06808	69.78539	2053388
438	191 844	20.92845	66.18157	2283105	488	238 144	22.09072	69.85700	2049180
439	192 721	20.95233	66.25708	2277904	489	239 121	22.11334	69.92853	2044990
440	193 600	20.97618	66.33250	2272727	490	240 100	22.13594	70.00000	2040816
441	194 481	21.00000	66.40783	2267574	491	241 081	22.15852	70.07139	2036660
442	195 364	21.02380	66.48308	2262443	492	242 064	22.18107	70.14271	2032520
443	196 249	21.04757	66.55825	2257336	493	243 049	22.20360	70.21396	2028398
444	197 136	21.07131	66.63332	2252252	494	244 036	22.22611	70.28513	2024291
445	198 025	21.09502	66.70832	2247191	495	245 025	22.24860	70.35624	2020202
446	198 916	21.11871	66.78323	2242152	496	246 016	22.27106	70.42727	2016129
447	199 809	21.14237	66.85806	2237136	497	247 009	22.29350	70.49823	2012072
448	200 704	21.16601	66.93280	2232143	498	248 004	22.31591	70.56912	2008032
449	201 601	21.18962	67.00746	2227171	499	249 001	22.33831	70.63993	2004008
450	202 500	21.21320	67.08204	2222222	500	250 000	22.36068	70.71068	2000000

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 500-600

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00	N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00
500	250 000	22.36068	70.71068	2000000	550	302 500	23.45208	74.16198	1818182
501	251 001	22.38303	70.78135	1996008	551	303 601	23.47339	74.22937	1814882
502	252 004	22.40536	70.85196	1992032	552	304 704	23.49468	74.29670	1811594
503	253 009	22.42766	70.92249	1988072	553	305 809	23.51595	74.36397	1808318
504	254 016	22.44994	70.99296	1984127	554	306 916	23.53720	74.43118	1805054
505	255 025	22.47221	71.06335	1980198	555	308 025	23.55844	74.49832	1801802
506	256 036	22.49444	71.13368	1976285	556	309 136	23.57965	74.56541	1798561
507	257 049	22.51666	71.20393	1972387	557	310 249	23.60085	74.63243	1795332
508	258 064	22.53886	71.27412	1968504	558	311 364	23.62202	74.69940	1792115
509	259 081	22.56103	71.34424	1964637	559	312 481	23.64318	74.76630	1788909
510	260 100	22.58318	71.41428	1960784	560	313 600	23.66432	74.83315	1785714
511	261 121	22.60531	71.48426	1956947	561	314 721	23.68544	74.89993	1782531
512	262 144	22.62742	71.55418	1953125	562	315 844	23.70654	74.96666	1779359
513	263 169	22.64950	71.62402	1949318	563	316 969	23.72762	75.03333	1776199
514	264 196	22.67157	71.69379	1945525	564	318 096	23.74868	75.09993	1773050
515	265 225	22.69361	71.76350	1941748	565	319 225	23.76973	75.16648	1769912
516	266 256	22.71563	71.83314	1937984	566	320 356	23.79075	75.23297	1766784
517	267 289	22.73763	71.90271	1934236	567	321 489	23.81176	75.29940	1763668
518	268 324	22.75961	71.97222	1930502	568	322 624	23.83275	75.36577	1760563
519	269 361	22.78157	72.04165	1926782	569	323 761	23.85372	75.43209	1757469
520	270 400	22.80351	72.11103	1923077	570	324 900	23.87467	75.49834	1754386
521	271 441	22.82542	72.18033	1919386	571	326 041	23.89561	75.56454	1751313
522	272 484	22.84732	72.24957	1915709	572	327 184	23.91652	75.63068	1748252
523	273 529	22.86919	72.31874	1912046	573	328 329	23.93742	75.69676	1745201
524	274 576	22.89105	72.38784	1908397	574	329 476	23.95830	75.76279	1742160
525	275 625	22.91288	72.45688	1904762	575	330 625	23.97916	75.82875	1739130
526	276 676	22.93469	72.52586	1901141	576	331 776	24.00000	75.89466	1736111
527	277 729	22.95648	72.59477	1897533	577	332 929	24.02082	75.96052	1733102
528	278 784	22.97825	72.66361	1893939	578	334 084	24.04163	76.02631	1730104
529	279 841	23.00000	72.73239	1890359	579	335 241	24.06242	76.09205	1727116
530	280 900	23.02173	72.80110	1886792	580	336 400	24.08319	76.15773	1724138
531	281 961	23.04344	72.86975	1883239	581	337 561	24.10394	76.22336	1721170
532	283 024	23.06513	72.93833	1879699	582	338 724	24.12468	76.28892	1718213
533	284 089	23.08679	73.00685	1876173	583	339 889	24.14539	76.35444	1715266
534	285 156	23.10844	73.07530	1872659	584	341 056	24.16609	76.41989	1712329
535	286 225	23.13007	73.14369	1869159	585	342 225	24.18677	76.48529	1709402
536	287 296	23.15167	73.21202	1865672	586	343 396	24.20744	76.55064	1706485
537	288 369	23.17326	73.28028	1862197	587	344 569	24.22808	76.61593	1703578
538	289 444	23.19483	73.34848	1858736	588	345 744	24.24871	76.68116	1700680
539	290 521	23.21637	73.41662	1855288	589	346 921	24.26932	76.74634	1697793
540	291 600	23.23790	73.48469	1851852	590	348 100	24.28992	76.81146	1694915
541	292 681	23.25941	73.55270	1848429	591	349 281	24.31049	76.87652	1692047
542	293 764	23.28089	73.62065	1845018	592	350 464	24.33105	76.94154	1689189
543	294 849	23.30236	73.68853	1841621	593	351 649	24.35159	77.00649	1686341
544	295 936	23.32381	73.75636	1838235	594	352 836	24.37212	77.07140	1683502
545	297 025	23.34524	73.82412	1834862	595	354 025	24.39262	77.13624	1680672
546	298 116	23.36664	73.89181	1831502	596	355 216	24.41311	77.20104	1677852
547	299 209	23.38803	73.95945	1828154	597	356 409	24.43358	77.26578	1675042
548	300 304	23.40940	74.02702	1824818	598	357 604	24.45404	77.33046	1672241
549	301 401	23.43075	74.09453	1821494	599	358 801	24.47448	77.39509	1669449
550	302 500	23.45208	74.16198	1818182	600	360 000	24.49490	77.45967	1666667

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 600-700

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00	N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00
600	360 000	24.49490	77.45967	1666667	650	422 500	25.49510	80.62258	1538462
601	361 201	24.51530	77.52419	1663894	651	423 801	25.51470	80.68457	1536098
602	362 404	24.53569	77.58866	1661130	652	425 104	25.53429	80.74652	1533742
603	363 609	24.55606	77.65307	1658375	653	426 409	25.55386	80.80842	1531394
604	364 816	24.57641	77.71744	1655629	654	427 716	25.57342	80.87027	1529052
605	366 025	24.59675	77.78175	1652893	655	429 025	25.59297	80.93207	1526718
606	367 236	24.61707	77.84600	1650165	656	430 336	25.61250	80.99383	1524390
607	368 449	24.63737	77.91020	1647446	657	431 649	25.63201	81.05554	1522070
608	369 664	24.65766	77.97435	1644737	658	432 964	25.65151	81.11720	1519757
609	370 881	24.67793	78.03845	1642036	659	434 281	25.67100	81.17881	1517451
610	372 100	24.69818	78.10250	1639344	660	435 600	25.69047	81.24038	1515152
611	373 321	24.71841	78.16649	1636661	661	436 921	25.70992	81.30191	1512859
612	374 544	24.73863	78.23043	1633987	662	438 244	25.72936	81.36338	1510574
613	375 769	24.75884	78.29432	1631321	663	439 569	25.74879	81.42481	1508296
614	376 996	24.77902	78.35815	1628664	664	440 896	25.76820	81.48620	1506024
615	378 225	24.79919	78.42194	1626016	665	442 225	25.78759	81.54753	1503759
616	379 456	24.81935	78.48567	1623377	666	443 556	25.80698	81.60882	1501502
617	380 689	24.83948	78.54935	1620746	667	444 889	25.82634	81.67007	1499250
618	381 924	24.85961	78.61298	1618123	668	446 224	25.84570	81.73127	1497006
619	383 161	24.87971	78.67655	1615509	669	447 561	25.86503	81.79242	1494768
620	384 400	24.89980	78.74008	1612903	670	448 900	25.88436	81.85353	1492537
621	385 641	24.91987	78.80355	1610306	671	450 241	25.90367	81.91459	1490313
622	386 884	24.93993	78.86698	1607717	672	451 584	25.92296	81.97561	1488095
623	388 129	24.95997	78.93035	1605136	673	452 929	25.94224	82.03658	1485884
624	389 376	24.97999	78.99367	1602564	674	454 276	25.96151	82.09750	1483680
625	390 625	25.00000	79.05694	1600000	675	455 625	25.98076	82.15838	1481481
626	391 876	25.01999	79.12016	1597444	676	456 976	26.00000	82.21922	1479290
627	393 129	25.03997	79.18333	1594896	677	458 329	26.01922	82.28001	1477105
628	394 384	25.05993	79.24645	1592357	678	459 684	26.03843	82.34076	1474926
629	395 641	25.07987	79.30952	1589825	679	461 041	26.05763	82.40146	1472754
630	396 900	25.09980	79.37254	1587302	680	462 400	26.07681	82.46211	1470588
631	398 161	25.11971	79.43551	1584786	681	463 761	26.09598	82.42272	1468429
632	399 424	25.13961	79.49843	1582278	682	465 124	26.11513	82.58329	1466276
633	400 689	25.15949	79.56130	1579779	683	466 489	26.13427	82.64381	1464129
634	401 956	25.17936	79.62412	1577287	684	467 856	26.15339	82.70429	1461988
635	403 225	25.19921	79.68689	1574803	685	469 225	26.17250	82.76473	1459854
636	404 496	25.21904	79.74961	1572327	686	470 596	26.19160	82.82512	1457726
637	405 769	25.23886	79.81228	1569859	687	471 969	26.21068	82.88546	1455604
638	407 044	25.25866	79.87490	1567398	688	473 344	26.22975	82.94577	1453488
639	408 321	25.27845	79.93748	1564945	689	474 721	26.24881	83.00602	1451379
640	409 600	25.29822	80.00000	1562500	690	476 100	26.26785	83.06624	1449275
641	410 881	25.31798	80.06248	1560062	691	477 481	26.28688	83.12641	1447178
642	412 164	25.33772	80.12490	1557632	692	478 864	26.30589	83.18654	1445087
643	413 449	25.35744	80.18728	1555210	693	480 249	26.32489	83.24662	1443001
644	414 736	25.37716	80.24961	1552795	694	481 636	26.34388	83.30666	1440922
645	416 025	25.39685	80.31189	1550388	695	483 025	26.36285	83.36666	1438849
646	417 316	25.41653	80.37413	1547988	696	484 416	26.38181	83.42661	1436782
647	418 609	25.43619	80.43631	1545595	697	485 809	26.40076	83.48653	1434720
648	419 904	25.45584	80.49845	1543210	698	487 204	26.41969	83.54639	1432665
649	421 201	25.47548	80.56054	1540832	699	488 601	26.43861	83.60622	1430615
650	422 500	25.49510	80.62258	1538462	700	490 000	26.45751	83.66600	1428571

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 700-800

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00	N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00
700	490 000	26.45751	83.66600	1428571	750	562 500	27.38613	86.60254	1333333
701	491 401	26.47640	83.72574	1426534	751	564 001	27.40438	86.66026	1331558
702	492 804	26.49528	83.78544	1424501	752	565 504	27.42262	86.71793	1329787
703	494 209	26.51415	83.84510	1422475	753	567 009	27.44085	86.77557	1328021
704	495 616	26.53300	83.90471	1420455	754	568 516	27.45906	86.83317	1326260
705	497 025	26.55184	83.96428	1418440	755	570 025	27.47726	86.89074	1324503
706	498 436	26.57066	84.02381	1416431	756	571 536	27.49545	86.94826	1322751
707	499 849	26.58947	84.08329	1414427	757	573 049	27.51363	87.00575	1321004
708	501 264	26.60827	84.14274	1412429	758	574 564	27.53180	87.06320	1319261
709	502 681	26.62705	84.20214	1410437	759	576 081	27.54995	87.12061	1317523
710	504 100	26.64583	84.26150	1408451	760	577 600	27.56810	87.17798	1315789
711	505 521	26.66458	84.32082	1406470	761	579 121	27.58623	87.23531	1314060
712	506 944	26.68333	84.38009	1404494	762	580 644	27.60435	87.29261	1312336
713	508 369	26.70206	84.43933	1402525	763	582 169	27.62245	87.34987	1310616
714	509 796	26.72078	84.49852	1400560	764	583 696	27.64055	87.40709	1308901
715	511 225	26.73948	84.55767	1398601	765	585 225	27.65863	87.46428	1307190
716	512 656	26.75818	84.61678	1396648	766	586 756	27.67671	87.52143	1305483
717	514 089	26.77686	84.67585	1394700	767	588 289	27.69476	87.57854	1303781
718	515 524	26.79552	84.73488	1392758	768	589 824	27.71281	87.63561	1302083
719	516 961	26.81418	84.79387	1390821	769	591 361	27.73085	87.69265	1300390
720	518 400	26.83282	84.85281	1388889	770	592 900	27.74887	87.74964	1298701
721	519 841	26.85144	84.91172	1386963	771	594 441	27.76689	87.80661	1297017
722	521 284	26.87006	84.97058	1385042	772	595 984	27.78489	87.86353	1295337
723	522 729	26.88866	85.02941	1383126	773	597 529	27.80288	87.92042	1293661
724	524 176	26.90725	85.08819	1381215	774	599 076	27.82086	87.97727	1291990
725	525 625	26.92582	85.14693	1379310	775	600 625	27.83882	88.03408	1290323
726	527 076	26.94439	85.20563	1377410	776	602 176	27.85678	88.09086	1288660
727	528 529	26.96294	85.26429	1375516	777	603 729	27.87472	88.14760	1287001
728	529 984	26.98148	85.32292	1373626	778	605 284	27.89265	88.20431	1285347
729	531 441	27.00000	85.38150	1371742	779	606 841	27.91057	88.26098	1283697
730	532 900	27.01851	85.44004	1369863	780	608 400	27.92848	88.31761	1282051
731	534 361	27.03701	85.49854	1367989	781	609 961	27.94638	88.37420	1280410
732	535 824	27.05550	85.55700	1366120	782	611 524	27.96426	88.43076	1278772
733	537 289	27.07397	85.61542	1364256	783	613 089	27.98214	88.48729	1277139
734	538 756	27.09243	85.67380	1362398	784	614 656	28.00000	88.54377	1275510
735	540 225	27.11088	85.73214	1360544	785	616 225	28.01785	88.60023	1273885
736	541 696	27.12932	85.79044	1358696	786	617 796	28.03569	88.65664	1272265
737	543 169	27.14774	85.84870	1356852	787	619 369	28.05352	88.71302	1270648
738	544 644	27.16616	85.90693	1355014	788	620 944	28.07134	88.76936	1269036
739	546 121	27.18455	85.96511	1353180	789	622 521	28.08914	88.82567	1267427
740	547 600	27.20294	86.02325	1351351	790	624 100	28.10694	88.88194	1265823
741	549 081	27.22132	86.08136	1349528	791	625 681	28.12472	88.93818	1264223
742	550 564	27.23968	86.13942	1347709	792	627 264	28.14249	88.99438	1262626
743	552 049	27.25803	86.19745	1345895	793	628 849	28.16026	89.05055	1261034
744	553 536	27.27636	86.25543	1344086	794	630 436	28.17801	89.10668	1259446
745	555 025	27.29469	86.31338	1342282	795	632 025	28.19574	89.16277	1257862
746	556 516	27.31300	86.37129	1340483	796	633 616	28.21347	89.21883	1256281
747	558 009	27.33130	86.42916	1338688	797	635 209	28.23119	89.27486	1254705
748	559 504	27.34959	86.48699	1336898	798	636 804	28.24889	89.33085	1253133
749	561 001	27.36786	86.54479	1335113	799	638 401	28.26659	89.38680	1251564
750	562 500	27.38613	86.60254	1333333	800	640 000	28.28427	89.44272	1250000

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 800-900

N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00	N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00
800	640 000	28.28427	89.44272	1250000	850	722 500	29.15476	92.19544	1176471
801	641 601	28.30194	89.49860	1248439	851	724 201	29.17190	92.24966	1175088
802	643 204	28.31960	89.55445	1246883	852	725 904	29.18904	92.30385	1173709
803	644 809	28.33725	89.61027	1245330	853	727 609	29.20616	92.35800	1172333
804	646 416	28.35489	89.66605	1243781	854	729 316	29.22328	92.41212	1170960
805	648 025	28.37252	89.72179	1242236	855	731 025	29.24038	92.46621	1169591
806	649 636	28.39014	89.77750	1240695	856	732 736	29.25748	92.52027	1168224
807	651 249	28.40775	89.83318	1239157	857	734 449	29.27456	92.57429	1166861
808	652 864	28.42533	89.88882	1237624	858	736 164	29.29164	92.62829	1165501
809	654 481	28.44293	89.94443	1236094	859	737 881	29.30870	92.68225	1164144
810	656 100	28.46050	90.00000	1234568	860	739 600	29.32576	92.73618	1162791
811	657 721	28.47806	90.05554	1233046	861	741 321	29.34280	92.79009	1161440
812	659 344	28.49561	90.11104	1231527	862	743 044	29.35984	92.84396	1160093
813	660 969	28.51315	90.16651	1230012	863	744 769	29.37686	92.89779	1158749
814	662 596	28.53069	90.22195	1228501	864	746 496	29.39388	92.95160	1157407
815	664 225	28.54820	90.27735	1226994	865	748 225	29.41088	93.00538	1156060
816	665 856	28.56571	90.33272	1225490	866	749 956	29.42788	93.05912	1154734
817	667 489	28.58321	90.38805	1223990	867	751 689	29.44486	93.11283	1153403
818	669 124	28.60070	90.44335	1222494	868	753 424	29.46184	93.16652	1152074
819	670 761	28.61818	90.49862	1221001	869	755 161	29.47881	93.22017	1150748
820	672 400	28.63564	90.55385	1219512	870	756 900	29.49576	93.27379	1149425
821	674 041	28.65310	90.60905	1218027	871	758 641	29.51271	93.32738	1148106
822	675 684	28.67058	90.66422	1216545	872	760 384	29.52965	93.38094	1146789
823	677 329	28.68798	90.71935	1215067	873	762 129	29.54657	93.43447	1145475
824	678 976	28.70540	90.77445	1213592	874	763 876	29.56349	93.48797	1144165
825	680 625	28.72281	90.82951	1212121	875	765 625	29.58040	93.54143	1142857
826	682 276	28.74022	90.88454	1210654	876	767 376	29.59730	93.59487	1141553
827	683 929	28.75761	90.93954	1209190	877	769 129	29.61419	93.64828	1140251
828	685 584	28.77499	90.99451	1207729	878	770 884	29.63106	93.70165	1138952
829	687 241	28.79236	91.04944	1206273	879	772 641	29.64793	93.75500	1137656
830	688 900	28.80972	91.10434	1204819	880	774 400	29.66479	93.80832	1136364
831	690 561	28.82707	91.15920	1203369	881	776 161	29.68164	93.86160	1135074
832	692 224	28.84441	91.21403	1201923	882	777 924	29.69848	93.91486	1133787
833	693 889	28.86174	91.26883	1200480	883	779 689	29.71532	93.96808	1132503
834	695 556	28.87906	91.32360	1199041	884	781 456	29.73214	94.02127	1131222
835	697 225	28.89637	91.37833	1197605	885	783 225	29.74895	94.07444	1129944
836	698 896	28.91366	91.43304	1196172	886	784 996	29.76575	94.12757	1128668
837	700 569	28.93095	91.48770	1194743	887	786 769	29.78255	94.18068	1127396
838	702 244	28.94823	91.54234	1193317	888	788 544	29.79933	94.23375	1126126
839	703 921	28.96550	91.59694	1191895	889	790 321	29.81610	94.28680	1124859
840	705 600	28.98275	91.65151	1190476	890	792 100	29.83287	94.33981	1123596
841	707 281	29.00000	91.70605	1189061	891	793 881	29.84962	94.39280	1122334
842	708 964	29.01724	91.76056	1187648	892	795 664	29.86637	94.44575	1121076
843	710 649	29.03446	91.81503	1186240	893	797 449	29.88311	94.49868	1119821
844	712 336	29.05168	91.86947	1184834	894	799 236	29.89983	94.55157	1118568
845	714 025	29.06888	91.92388	1183432	895	801 025	29.91655	94.60444	1117318
846	715 716	29.08608	91.97826	1182033	896	802 816	29.93326	94.65728	1116071
847	717 409	29.10326	92.03260	1180638	897	804 609	29.94996	94.71008	1114827
848	719 104	29.12044	92.08692	1179245	898	806 404	29.96665	94.76286	1113586
849	720 801	29.13760	92.14120	1177856	899	808 201	29.98333	94.81561	1112347
850	722 500	29.15476	92.19544	1176471	900	810 000	30.00000	94.86833	1111111

SQUARES, SQUARE ROOTS, AND RECIPROCAL: 900-1000

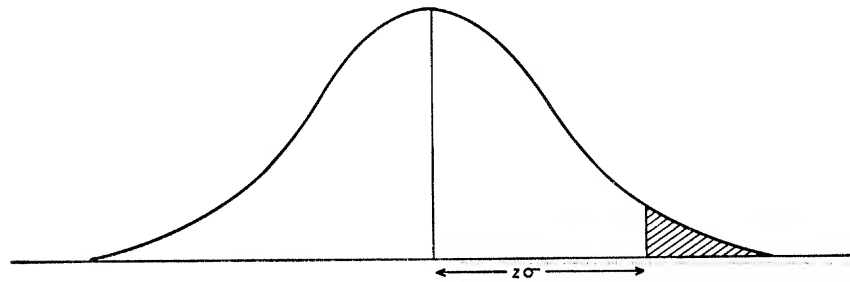
N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00	N	N^2	$\sqrt{N}$	$\sqrt{10N}$	$1/N$.00
900	810 000	30.00000	94.86833	1111111	950	902 500	30.82207	97.46794	1052632
901	811 801	30.01666	94.92102	1109878	951	904 401	30.83829	97.51923	1051525
902	813 604	30.03331	94.97368	1108647	952	906 304	30.85450	97.57049	1050420
903	815 409	30.04996	95.02631	1107420	953	908 209	30.87070	97.62172	1049318
904	817 216	30.06659	95.07891	1106195	954	910 116	30.88689	97.67292	1048218
905	819 025	30.08322	95.13149	1104972	955	912 025	30.90307	97.72410	1047120
906	820 836	30.09983	95.18403	1103753	956	913 936	30.91925	97.77525	1046025
907	822 649	30.11644	95.23655	1102536	957	915 849	30.93542	97.82638	1044932
908	824 464	30.13304	95.28903	1101322	958	917 764	30.95158	97.87747	1043841
909	826 281	30.14963	95.34149	1100110	959	919 681	30.96773	97.92855	1042753
910	828 100	30.16621	95.39392	1098901	960	921 600	30.98387	97.97959	1041667
911	829 921	30.18278	95.44632	1097695	961	923 521	31.00000	98.03061	1040583
912	831 744	30.19934	95.49869	1096491	962	925 444	31.01612	98.08160	1039501
913	833 569	30.21589	95.55103	1095290	963	927 369	31.03224	98.13256	1038422
914	835 396	30.23243	95.60335	1094092	964	929 296	31.04835	98.18350	1037344
915	837 225	30.24897	95.65563	1092896	965	931 225	31.06445	98.23441	1036269
916	839 056	30.26549	95.70789	1091703	966	933 156	31.08054	98.28530	1035197
917	840 889	30.28201	95.76012	1090513	967	935 089	31.09662	98.33616	1034126
918	842 724	30.29851	95.81232	1089325	968	937 024	31.11270	98.38699	1033058
919	844 561	30.31501	95.86449	1088139	969	938 961	31.12876	98.43780	1031992
920	846 400	30.33150	95.91663	1086957	970	940 900	31.14482	98.48858	1030928
921	848 241	30.34798	95.96874	1085776	971	942 841	31.16087	98.53933	1029866
922	850 084	30.36445	96.02083	1084599	972	944 784	31.17691	98.59006	1028807
923	851 929	30.38092	96.07289	1083424	973	946 729	31.19295	98.64076	1027749
924	853 776	30.39737	96.12492	1082251	974	948 676	31.20897	98.69144	1026694
925	855 625	30.41381	96.17692	1081081	975	950 625	31.22499	98.74209	1025641
926	857 476	30.43025	96.22889	1079914	976	952 576	31.24100	98.79271	1024590
927	859 329	30.44667	96.28084	1078749	977	954 529	31.25700	98.84331	1023541
928	861 184	30.46309	96.33276	1077586	978	956 484	31.27299	98.89388	1022495
929	863 041	30.47950	96.38465	1076426	979	958 441	31.28898	98.94443	1021450
930	864 900	30.49590	96.43651	1075269	980	960 400	31.30495	98.99495	1020408
931	866 761	30.51229	96.48834	1074114	981	962 361	31.32092	99.04544	1019368
932	868 624	30.52868	96.54015	1072961	982	964 324	31.33688	99.09591	1018330
933	870 489	30.54505	96.59193	1071811	983	966 289	31.35283	99.14636	1017294
934	872 356	30.56141	96.64368	1070664	984	968 256	31.36877	99.19677	1016260
935	874 225	30.57777	96.69540	1069519	985	970 225	31.38471	99.24717	1015228
936	876 096	30.59412	96.74709	1068376	986	972 196	31.40064	99.29753	1014199
937	877 969	30.61046	96.79876	1067236	987	974 169	31.41656	99.34787	1013171
938	879 844	30.62679	96.85040	1066098	988	976 144	31.43247	99.39819	1012146
939	881 721	30.64311	96.90201	1064963	989	978 121	31.44837	99.44848	1011122
940	883 600	30.65942	96.95360	1063830	990	980 100	31.46427	99.49874	1010101
941	885 481	30.67572	97.00515	1062699	991	982 081	31.48015	99.54898	1009082
942	887 364	30.69202	97.05668	1061571	992	984 064	31.49603	99.59920	1008065
943	889 249	30.70831	97.10819	1060445	993	986 049	31.51190	99.64939	1007049
944	891 136	30.72458	97.15966	1059322	994	988 036	31.52777	99.69955	1006036
945	893 025	30.74085	97.21111	1058201	995	990 025	31.54362	99.74969	1005025
946	894 916	30.75711	97.26253	1057082	996	992 016	31.55947	99.79980	1004016
947	896 809	30.77337	97.31393	1055966	997	994 009	31.57531	99.84989	1003009
948	898 704	30.78961	97.36529	1054852	998	996 004	31.59114	99.89995	1002004
949	900 601	30.80584	97.41663	1053741	999	998 001	31.60696	99.94999	1001001
950	902 500	30.82207	97.46794	1052632	1000	1 000 000	31.62278	100.00000	1000000

APPENDIX • B

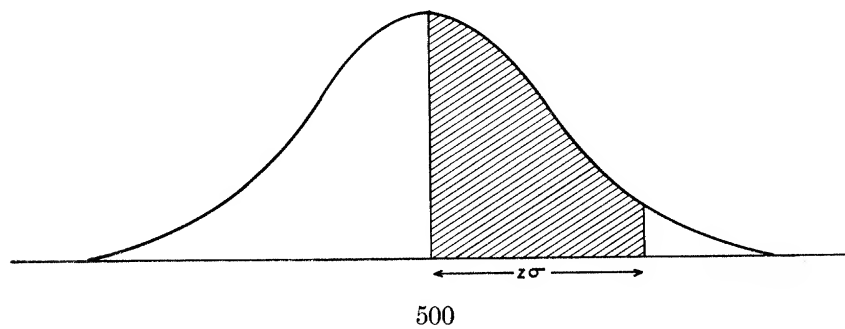
Areas under the Normal Curve and Normal Deviates

The two tables given on the following pages are used to measure areas (relative frequencies or probabilities) under the normal curve.

Table B.1 shows, for the value of a given normal deviate (z), the area in the tail of a normal curve; that is, it indicates the area depicted below:



To determine the area in a central portion of the normal curve—namely, the following shaded area:



we simply subtract the appropriate value shown in Table B.1 from .5000.

Table B.2 shows, for a given tail area, the value of the corresponding normal deviate.

TABLE B.1. AREAS UNDER THE NORMAL CURVE

Normal De- viate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

TABLE B.2. NORMAL DEVIATES

Single-Tail Area	Two-Tail Area	z	Single-Tail Area	Two-Tail Area	z
.0025	.005	2.81	.06	.12	1.55
.005	.010	2.58	.07	.14	1.48
.010	.020	2.33	.08	.16	1.41
.0125	.025	2.24	.09	.18	1.34
.015	.030	2.17	.10	.20	1.28
.020	.040	2.05	.20	.40	.84
.025	.050	1.96	.25	.50	.67
.030	.060	1.88	.30	.60	.52
.040	.080	1.75	.40	.80	.25
.050	.100	1.64	.50	1.00	0

APPENDIX • C

Table of 5,000 Random Digits*

A table of 5,000 random digits is reproduced on the following pages. As indicated in Section 4.7, random digits are useful to draw probability samples. The digits given come from a much larger table compiled for use in sampling studies by The Rand Corporation.

RANDOM DIGITS

Line No.	Column Number									
	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
0	93108	77033	68325	10160	38667	62441	87023	94372	06164	30700
1	28271	08589	83279	48838	60935	70541	53814	95588	05832	80235
2	21841	35545	11148	34775	17308	88034	97765	35959	52843	44895
3	22025	79554	19698	25255	50283	94037	57463	92925	12042	91414
4	09210	20779	02994	02258	86978	85092	54052	18354	20914	28460
5	90552	71129	03621	20517	16908	06668	29916	51537	93658	29525
6	01130	06995	20258	10351	99248	51660	38861	49668	74742	47181
7	22604	56719	21784	68788	38358	59827	19270	99287	81193	43366
8	06690	01800	34272	65497	94891	14537	91358	21587	95765	72605
9	59809	69982	71809	64984	48709	43991	24987	69246	86400	29559
10	56475	02726	58511	95405	70293	84971	06676	44075	32338	31980
11	02730	34870	83209	03138	07715	31557	55242	61308	26507	06186
12	74482	33990	13509	92588	10462	76546	46097	01825	20153	36271
13	19793	22487	94238	81054	95488	23617	15539	94335	73822	93481
14	19020	27856	60526	24144	98021	60564	46373	86928	52135	74919
15	69565	60635	65709	77887	42766	86698	14004	94577	27936	47220
16	69274	23208	61035	84263	15034	28717	76146	22021	23779	98562
17	83658	14204	09445	41081	49630	34215	89806	40930	97194	21747
18	78612	51102	66826	40430	54072	62164	68977	95583	11765	81072
19	14980	74158	78216	38985	60838	82836	42777	85321	90463	11813

* From The Rand Corporation, *A Million Random Digits* (Glencoe, Ill.: The Free Press, 1955), by permission.

RANDOM DIGITS

Line No.	Column Number									
	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
20	63172	28010	29405	91554	75195	51183	65805	87525	35952	83204
21	71167	37984	52737	06869	38122	95322	41356	19391	96787	64410
22	78530	56410	19195	34434	83712	50397	80920	15464	81350	18673
23	98324	03774	07573	67864	06497	20758	83454	22756	83959	96347
24	55793	30055	08373	32652	02654	75980	02095	87545	88815	80086
25	05674	34471	61967	91266	38814	44728	32455	17057	08339	93997
26	15643	22245	07592	22078	73628	60902	41561	54608	41023	98345
27	66750	19609	70358	03622	64898	82220	69304	46235	97332	64539
28	42320	74314	50222	82339	51564	42885	50482	98501	02245	88990
29	73752	73818	15470	04914	24936	65514	56633	72030	30856	85183
30	97546	02188	46373	21486	28221	08155	23486	66134	88799	49496
31	32569	52162	38444	42004	78011	16909	94194	79732	47114	23919
32	36048	93973	82596	28739	86985	58144	65007	08786	14826	04896
33	40455	36702	38965	56042	80023	28169	04174	65533	52718	55255
34	33597	47071	55618	51796	71027	46690	08002	45066	02870	60012
35	22828	96380	35883	15910	17211	42358	14056	55438	98148	35384
36	00631	95925	19324	31497	88118	06283	84596	72091	53987	01477
37	75722	36478	07634	63114	27164	15467	03983	09141	60562	65725
38	80577	01771	61510	17099	28731	41426	18853	41523	14914	76661
39	10524	20900	65463	83680	05005	11611	64426	59065	06758	02892
40	93815	69446	75253	51915	97839	75427	90685	60352	96288	34248
41	81867	97119	93446	20862	46591	97677	42704	13718	44975	67145
42	64649	07689	16711	12169	15238	74106	60655	56289	74166	78561
43	55768	09210	52439	33355	57884	36791	00853	49969	74814	09270
44	38080	49460	48137	61589	42742	92035	21766	19435	92579	27683
45	22360	16332	05343	34613	24013	98831	17157	44089	07366	66196
46	40521	09057	00239	51284	71556	22605	41293	54854	39736	05113
47	19292	69862	59951	49644	53486	28244	20714	56030	39292	45166
48	79504	40078	06838	05509	68581	39400	85615	52314	83202	40313
49	64138	27983	84048	42631	58658	62243	82572	45211	37060	15017
50	31000	63837	17813	08076	19164	95508	17513	29416	61238	25818
51	87014	83331	56364	32768	85680	08844	59844	68794	32783	60318
52	47293	97023	35804	69886	47494	94574	45842	67221	77115	43398
53	33898	89236	06100	68848	08674	87786	42425	92091	86274	82166
54	26877	95856	65227	25165	01752	99463	15216	28719	04716	80246
55	53984	87855	70753	80386	78600	39244	76967	83263	57849	85890
56	97809	03548	00574	21143	11605	30245	87395	80966	28721	11095
57	85683	79483	74858	87491	57785	61270	51111	50490	40940	02832
58	07758	57784	22934	17165	37776	33361	79191	97398	40881	98552
59	13995	74006	90843	85761	89037	13567	67089	47435	75156	70217
60	89630	94575	64517	80897	04861	50564	15287	94279	69154	75473
61	20355	08661	11092	19682	84287	23597	48246	74325	66320	90155
62	13496	74367	32701	87819	29050	90959	99765	59374	45204	87750
63	68419	49317	30340	62744	45214	79826	34043	73218	72788	26143
64	61481	09786	08768	93715	43847	68901	51249	35382	67776	06271

RANDOM DIGITS

Line No.	Column Number									
	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
65	71877	53924	09630	62975	64470	46832	40591	23477	21063	31229
66	99288	86530	86819	38955	91744	89632	64623	84986	85990	47639
67	09047	83453	12775	32000	60098	98862	93422	66212	86413	82884
68	83724	89682	53950	41500	16023	83862	88274	68301	99673	40945
69	56126	46618	87708	91994	38384	27047	91550	18903	45535	45586
70	13810	51503	55122	22997	82556	22697	54985	64852	58775	66737
71	85630	22794	81623	10927	52252	27384	03122	17454	03250	23232
72	93350	28643	39097	00914	69844	06450	32818	88752	28722	94656
73	05002	71178	15414	99874	58046	35461	56349	81936	14964	83638
74	70358	57182	45747	30830	70542	25932	16298	23521	72454	11640
75	98169	51114	18115	40718	46082	75847	15678	22842	44615	97810
76	28010	95831	94028	57041	75137	55128	99785	33467	33834	04860
77	05576	47156	09300	11197	97670	20064	98145	84774	07907	10992
78	80748	66311	79421	00202	68501	40422	07368	26795	24358	78969
79	39746	23690	67845	83962	09451	21501	75317	09049	69440	16137
80	39380	72397	88106	07851	67756	58042	44218	52775	49082	54400
81	55865	70318	58189	35340	91500	80940	39231	54836	18038	03557
82	34223	72744	83926	40078	80791	42723	68340	47452	17443	69289
83	91312	44474	09925	41416	96671	11213	60979	04130	72380	73582
84	51010	08021	53914	37499	40228	92606	07225	18014	71063	50111
85	48890	58264	18790	91535	04933	56656	76389	37904	98017	07663
86	24876	93198	77444	53782	48866	65614	11410	90637	63675	43554
87	49408	42923	28162	09789	15155	66742	66785	79065	93573	19853
88	56307	90901	18118	36505	01598	68997	96338	40823	96247	55573
89	48320	09322	13457	84931	54701	14878	05422	65358	85636	31948
90	51083	00619	05548	25431	15175	82428	00637	41814	68871	31688
91	43249	71593	85595	59834	01488	06917	32858	80134	01832	25905
92	00370	17413	75537	79824	24428	28941	58659	66731	99940	27156
93	55737	45462	12484	64858	43581	06220	07507	39119	38024	99720
94	39551	38058	10445	18463	80812	51243	22351	63266	94057	06573
95	02369	11289	99499	32922	88429	90484	42010	53308	33206	36137
96	92788	62546	26147	83529	10012	77611	78925	86071	66344	35705
97	98650	95898	74254	45173	17430	73882	03411	88447	43279	35057
98	95591	85858	51058	26140	00995	16881	87372	72646	18796	58537
99	67853	23563	41063	63355	72454	16016	72229	54720	09846	81392

APPENDIX • D

Table of Common Logarithms, 100–999

The table on the following two pages gives the common logarithms of the numbers 100 through 999. Logarithms may be used to simplify the arithmetical operations of multiplication, division, raising numbers to powers, and extracting roots of numbers. Thus, the student in statistics may find them useful, for example, in the calculation of probabilities by the binomial formula. Calculations with the table of four-place logarithms given should yield results accurate to four digits.

For a discussion of the use of logarithms the student may consult almost any elementary college mathematics textbook. For example, see

COOLEY, HOLLIS R., GANS, DAVID, KLINE, MORRIS, and WAHLERT, HOWARD E. *Introduction to Mathematics*, pp. 123–136. 2nd ed. Boston: Houghton Mifflin Company, 1949.

COMMON LOGARITHMS: 100-549

N	0	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6336	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396

COMMON LOGARITHMS: 550-999

N	0	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996

Tables of the t Distribution and the F Distribution

Not all sampling distributions of pertinent statistics can be approximated by the normal curve. In some situations statisticians must resort to the use of some other mathematical curve to approximate a sampling distribution adequately. Two frequently useful curves or distributions of this sort are (i) the t distribution and (ii) the F distribution.

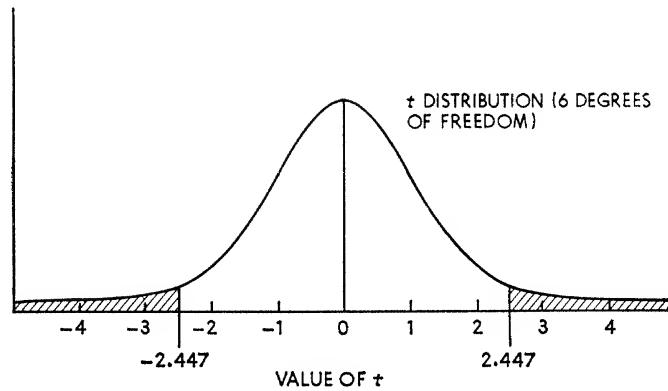
Just as the normal curve depends upon its mean and its standard deviation, the t distribution depends upon a parameter—generally called *degrees of freedom* (n). For one degree of freedom there is one t distribution, for two degrees of freedom there is another t distribution, for three degrees of freedom there is still another, and so forth ad infinitum.

For various t distributions (i.e., for various degrees of freedom: 1 to 30 and ∞) Table E.1 shows the value of t for a given area in both tails of the curve. This table, for t , is comparable to the table of normal deviates on page 502.

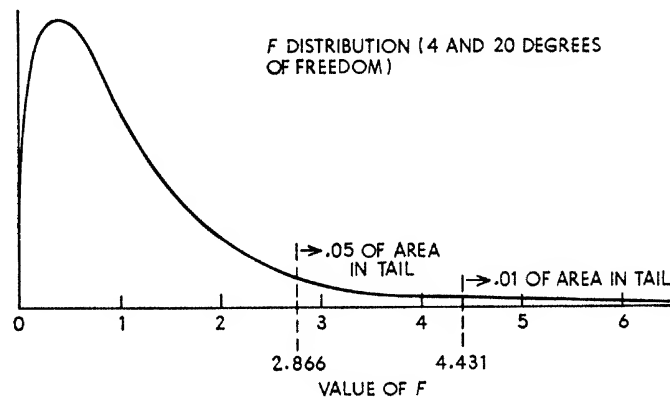
The graphic illustration of a t distribution on page 510 provides an example of the use of Table E.1.

Just as there are many t distributions, there are many F distributions—the shape of a particular F distribution depends upon a pair of degrees of freedom (n_1 and n_2). Table E.2 shows, for various pairs of degrees of freedom, the value of F for a given area under the curve's upper tail.

The graphic illustration of an F distribution on page 510 provides an example of the use of Table E.2.

GRAPHIC ILLUSTRATIONS OF THE t DISTRIBUTION AND THE F DISTRIBUTION

As Table E.1 indicates, for the t distribution with 6 degrees of freedom, .05 of the area in two tails of the curve falls outside of the range $t = -2.447$ to $t = 2.447$.



As Table E.2 indicates, for the F distribution with 4 and 20 degrees of freedom, .05 of the area under the curve is past $F = 2.866$ while .01 of the area is past $F = 4.431$.

TABLE E.1. VALUES OF t

n	.9	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.057
2	.142	.289	.445	.617	.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	.424	.584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750
∞	.12566	.25335	.38532	.52440	.67449	.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582

* Reprinted from Table IV, p. 174, of R. A. Fisher, *Statistical Methods for Research Workers* (11th ed.), published by Oliver and Boyd, Ltd., Edinburgh, by permission of the author and publishers.

TABLE E.2. VALUES OF F

n ₂	n ₁ = 1			n ₁ = 2			n ₁ = 3			n ₁ = 4			n ₁ = 5		
	.05	.01	.001	.05	.01	.001	.05	.01	.001	.05	.01	.001	.05	.01	.001
1	161.45	4.052	405.284	199.50	4.999	600.000	215.71	5.403	540.379	224.58	5.694	569.500	230.16	5.763	576.405
2	18.513	98.503	998.5	19.000	30.817	999.0	19.164	99.166	999.2	19.147	99.249	999.2	19.296	99.299	999.3
3	10.128	34.116	167.5	9.552	30.817	148.5	9.277	29.457	141.1	9.171	28.710	137.1	9.014	28.237	134.6
4	7.709	21.198	74.14	6.944	18.000	61.25	6.591	16.694	58.18	6.388	15.977	55.44	6.256	15.523	54.71
5	6.608	16.268	47.04	5.786	13.274	36.61	5.410	12.650	35.20	5.192	11.392	31.09	5.050	10.967	29.75
6	5.987	13.745	35.51	5.143	10.925	27.00	4.757	9.779	25.70	4.534	9.149	21.90	4.387	8.745	20.81
7	5.591	12.246	29.22	4.737	9.547	21.69	4.347	9.431	18.77	4.120	8.847	17.19	3.972	8.460	16.21
8	5.318	11.269	25.42	4.459	8.649	18.49	4.066	8.591	16.53	3.838	8.006	14.89	3.682	7.932	14.49
9	5.117	10.561	22.86	4.256	8.032	16.39	3.863	7.992	15.20	3.633	7.432	13.56	3.482	7.467	13.11
10	4.965	10.044	21.04	4.103	7.559	14.91	3.708	7.552	13.85	3.478	6.994	12.28	3.326	6.956	11.71
11	4.844	9.646	19.69	3.982	7.206	13.81	3.587	7.217	12.56	3.357	6.685	11.25	3.204	6.655	10.43
12	4.747	9.330	18.64	3.885	6.927	12.97	3.490	6.953	11.80	3.259	6.412	10.35	3.106	6.404	9.58
13	4.667	9.074	17.81	3.806	6.701	12.31	3.410	6.733	11.24	3.179	6.205	9.97	3.025	6.205	8.89
14	4.600	8.862	17.14	3.739	6.515	11.78	3.344	6.564	10.81	3.113	6.035	9.62	2.958	6.035	8.25
15	4.543	8.683	16.59	3.682	6.359	11.34	3.287	6.417	10.39	3.056	5.893	9.25	2.901	5.893	7.67
16	4.494	8.531	16.12	3.634	6.226	10.97	3.239	6.292	9.90	3.007	5.773	8.94	2.852	5.773	7.27
17	4.451	8.400	15.72	3.592	6.112	10.66	3.197	6.185	9.73	2.965	5.669	8.68	2.810	5.669	7.02
18	4.414	8.288	15.38	3.555	6.013	10.39	3.159	6.092	9.49	2.928	5.579	8.43	2.773	5.579	6.81
19	4.381	8.188	15.08	3.522	5.926	10.16	3.126	6.010	9.28	2.895	5.500	8.20	2.740	5.500	6.61
20	4.351	8.096	14.82	3.493	5.849	9.95	3.098	5.938	9.10	2.866	5.431	8.00	2.711	5.431	6.46
21	4.325	8.017	14.59	3.467	5.780	9.77	3.072	5.874	8.94	2.840	5.369	7.85	2.685	5.369	6.32
22	4.301	7.946	14.38	3.443	5.719	9.61	3.046	5.817	8.79	2.817	5.313	7.69	2.661	5.313	6.19
23	4.279	7.881	14.19	3.422	5.664	9.47	3.028	5.765	8.67	2.795	5.264	7.55	2.640	5.264	6.08
24	4.260	7.823	14.03	3.403	5.614	9.34	3.009	5.718	8.55	2.776	5.218	7.42	2.621	5.218	5.98
25	4.242	7.770	13.88	3.385	5.568	9.22	2.991	5.676	8.45	2.759	5.177	7.30	2.603	5.177	5.88
26	4.225	7.721	13.74	3.369	5.526	9.12	2.975	5.637	8.36	2.743	5.140	7.20	2.587	5.140	5.80
27	4.210	7.677	13.61	3.354	5.488	9.03	2.960	5.601	8.27	2.728	5.106	7.11	2.572	5.106	5.73
28	4.196	7.636	13.50	3.340	5.453	8.93	2.947	5.568	8.19	2.714	5.074	7.02	2.558	5.074	5.66
29	4.183	7.598	13.39	3.328	5.421	8.85	2.934	5.538	8.12	2.701	5.045	6.94	2.545	5.045	5.59
30	4.171	7.563	13.29	3.316	5.390	8.77	2.922	5.510	8.05	2.690	5.018	6.86	2.534	5.018	5.53
40	4.085	7.314	12.61	3.232	5.178	8.25	2.839	5.313	7.60	2.606	4.828	6.50	2.450	4.828	5.13
60	4.001	7.077	11.37	3.150	4.977	7.51	2.758	5.126	6.87	2.525	4.649	6.01	2.368	4.649	4.76
120	3.920	6.861	10.38	3.072	4.785	7.01	2.680	4.949	6.49	2.447	4.480	5.62	2.290	4.480	4.42
∞	3.841	6.638	10.831	2.996	4.605	6.91	2.605	4.752	6.43	2.372	4.312	5.48	2.214	4.312	4.10

m	n ₁ = 6				n ₁ = 8				n ₁ = 12				n ₁ = 24				n ₁ = ∞			
	.05	.01	.001	.0001	.05	.01	.001	.0001	.05	.01	.001	.0001	.05	.01	.001	.0001	.05	.01	.001	.0001
1	233.90	5.859	0	585.987	238.88	5.981	0	598.144	243.91	6.106	0	610.607	249.05	6.234	0	623.497	254.32	6.366	0	636.619
2	19.330	99.332	999.8	999.8	19.371	99.374	999.8	999.8	19.413	99.416	999.8	999.8	19.454	99.458	999.8	999.8	19.496	99.501	999.8	999.8
3	8.941	27.911	129.8	129.8	8.981	27.953	129.8	129.8	8.923	27.995	129.8	129.8	8.965	28.037	129.8	129.8	8.907	28.079	129.8	129.8
4	6.163	18.207	26.68	26.68	6.203	18.249	26.68	26.68	6.145	18.291	26.68	26.68	6.187	18.333	26.68	26.68	6.129	18.375	26.68	26.68
5	4.950	10.672	28.84	28.84	5.000	10.714	28.84	28.84	4.940	10.756	28.84	28.84	4.980	10.798	28.84	28.84	4.920	10.840	28.84	28.84
6	4.284	8.456	80.08	80.08	4.334	8.498	80.08	80.08	4.274	8.540	80.08	80.08	4.314	8.582	80.08	80.08	4.254	8.624	80.08	80.08
7	3.866	7.191	45.63	45.63	3.916	7.233	45.63	45.63	3.856	7.275	45.63	45.63	3.896	7.317	45.63	45.63	3.836	7.359	45.63	45.63
8	3.581	6.371	19.86	19.86	3.631	6.413	19.86	19.86	3.571	6.455	19.86	19.86	3.611	6.497	19.86	19.86	3.551	6.539	19.86	19.86
9	3.374	5.802	11.98	11.98	3.424	5.844	11.98	11.98	3.364	5.886	11.98	11.98	3.404	5.928	11.98	11.98	3.344	5.970	11.98	11.98
10	3.217	5.386	9.02	9.02	3.267	5.426	9.02	9.02	3.207	5.468	9.02	9.02	3.247	5.510	9.02	9.02	3.187	5.552	9.02	9.02
11	3.095	5.069	6.05	6.05	3.145	5.109	6.05	6.05	3.085	5.151	6.05	6.05	3.125	5.193	6.05	6.05	3.065	5.235	6.05	6.05
12	2.996	4.821	4.88	4.88	3.046	4.863	4.88	4.88	2.986	4.905	4.88	4.88	3.026	4.947	4.88	4.88	2.966	4.989	4.88	4.88
13	2.915	4.620	3.86	3.86	2.965	4.662	3.86	3.86	2.905	4.704	3.86	3.86	2.945	4.746	3.86	3.86	2.885	4.788	3.86	3.86
14	2.848	4.456	3.03	3.03	2.898	4.498	3.03	3.03	2.838	4.540	3.03	3.03	2.878	4.582	3.03	3.03	2.818	4.624	3.03	3.03
15	2.790	4.318	2.39	2.39	2.840	4.360	2.39	2.39	2.780	4.402	2.39	2.39	2.820	4.444	2.39	2.39	2.760	4.486	2.39	2.39
16	2.741	4.202	1.91	1.91	2.791	4.244	1.91	1.91	2.731	4.286	1.91	1.91	2.771	4.328	1.91	1.91	2.711	4.370	1.91	1.91
17	2.699	4.102	1.56	1.56	2.749	4.146	1.56	1.56	2.689	4.188	1.56	1.56	2.729	4.230	1.56	1.56	2.669	4.272	1.56	1.56
18	2.661	4.015	1.30	1.30	2.711	4.059	1.30	1.30	2.651	4.101	1.30	1.30	2.691	4.143	1.30	1.30	2.631	4.185	1.30	1.30
19	2.628	3.939	1.10	1.10	2.678	3.981	1.10	1.10	2.618	4.023	1.10	1.10	2.658	4.065	1.10	1.10	2.598	4.107	1.10	1.10
20	2.599	3.871	0.94	0.94	2.649	3.913	0.94	0.94	2.589	3.955	0.94	0.94	2.629	3.997	0.94	0.94	2.569	4.039	0.94	0.94
21	2.573	3.812	0.81	0.81	2.621	3.854	0.81	0.81	2.561	3.896	0.81	0.81	2.601	3.938	0.81	0.81	2.541	3.980	0.81	0.81
22	2.549	3.758	0.70	0.70	2.597	3.800	0.70	0.70	2.537	3.842	0.70	0.70	2.577	3.884	0.70	0.70	2.517	3.926	0.70	0.70
23	2.528	3.710	0.60	0.60	2.576	3.752	0.60	0.60	2.516	3.794	0.60	0.60	2.556	3.836	0.60	0.60	2.496	3.978	0.60	0.60
24	2.508	3.667	0.55	0.55	2.556	3.709	0.55	0.55	2.496	3.749	0.55	0.55	2.536	3.791	0.55	0.55	2.476	3.930	0.55	0.55
25	2.490	3.627	0.46	0.46	2.537	3.668	0.46	0.46	2.477	3.700	0.46	0.46	2.517	3.742	0.46	0.46	2.457	3.982	0.46	0.46
26	2.474	3.591	0.38	0.38	2.521	3.632	0.38	0.38	2.461	3.664	0.38	0.38	2.501	3.704	0.38	0.38	2.441	3.944	0.38	0.38
27	2.459	3.556	0.31	0.31	2.505	3.597	0.31	0.31	2.445	3.628	0.31	0.31	2.485	3.666	0.31	0.31	2.425	3.906	0.31	0.31
28	2.445	3.523	0.24	0.24	2.491	3.564	0.24	0.24	2.431	3.635	0.24	0.24	2.471	3.673	0.24	0.24	2.411	3.868	0.24	0.24
29	2.432	3.499	0.18	0.18	2.478	3.540	0.18	0.18	2.418	3.610	0.18	0.18	2.458	3.648	0.18	0.18	2.398	3.830	0.18	0.18
30	2.421	3.474	0.12	0.12	2.466	3.517	0.12	0.12	2.406	3.587	0.12	0.12	2.446	3.625	0.12	0.12	2.387	3.792	0.12	0.12
40	2.336	3.291	0.07	0.07	2.380	3.461	0.07	0.07	2.324	3.531	0.07	0.07	2.368	3.591	0.07	0.07	2.312	3.756	0.07	0.07
60	2.254	3.119	0.04	0.04	2.297	3.383	0.04	0.04	2.241	3.453	0.04	0.04	2.285	3.513	0.04	0.04	2.229	3.680	0.04	0.04
120	2.175	3.056	0.03	0.03	2.216	3.335	0.03	0.03	2.160	3.405	0.03	0.03	2.204	3.465	0.03	0.03	2.148	3.632	0.03	0.03
∞	2.099	3.002	0.02	0.02	2.138	3.291	0.02	0.02	2.082	3.361	0.02	0.02	2.126	3.421	0.02	0.02	2.070	3.588	0.02	0.02

Source: Frederick E. Croxton and Dudley J. Cowden, *Practical Business Statistics* (2d ed.; New York: Prentice-Hall, Inc., 1948), pp. 514 and 515. Reprinted by permission of the authors and the publisher.

Values of F at the .05 and .01 points were taken, by permission, from Maxine Merrington and Catherine M. Thompson, "New Tables of Statistical Variables," *Biometrika*, Vol. XXXIII, Part 1, pp. 80, 81, 84, and 85. Values of F at the .001 point were taken from Table V of R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, Ltd., Edinburgh, 1948, by permission of the author and publishers. The first reference gives F values to five digits at the .50, .25, .10, .05, .025, .01, and .005 points and for values of m in addition to those shown in the above table. The second reference gives F values to three or more digits at the .20, .05, .01, and .001 points.

Index

A

- A_2 , 346
- Absolute deviation, 70
- Absolute seasonal, 477
- Acceptable quality level (AQL), 258
- Acceptance number, 260, 264-65
- Acceptance sampling
 - acceptable quality level (AQL), 258
 - acceptance number, 260, 264-65
 - alternate management objectives, 269-71
 - alternate plans, 263-68
 - average sample size, 266-67
 - basic principles of, 261-62
 - bias in selection procedure, 282
 - consumer's risk, 258
 - decision rule, 259-61
 - defined, 257
 - double sampling plan, 263-64, 266-68
 - fixed sample size, 263
 - lot tolerance per cent defective (LTPD), 258
 - minimum inspection costs, 269-70
 - multiple sampling plan, 263-64, 266-68
 - producer's risk, 258
 - sequential sampling plan, 263, 265-68
 - single sampling plan, 263-68
 - truncated sampling plan, 263
- Accuracy, 276, 303-5
- Action, principle of, 3-6
- Aggregate
 - in analytical studies, 60
 - as a decision-parameter, 58-62
 - in enumerative studies, 60
 - estimator for, 199
 - formula for, 61-62
 - of a population (A), 58-62
- Alpha risk (α -risk), 235
- Alternate hypothesis, 238
- Alternative courses of action, 5-6, 13
- American Management Association, 481
- American Society for Quality Control, 368
- American Society for Testing Materials, 368
- American Standards Association, Inc., 47, 368

- Analysis of variance
 - assumptions, 388
 - described, 387-91
 - F distribution, 391, 509
 - fixed model, 391
 - random model, 391
 - variance ratio, 390-91
- Analytical study, 28-31, 48
- Applied Statistics*, 299
- Arithmetic mean
 - and the aggregate, 65-66
 - in analytical studies, 66
 - arithmetical properties, 66
 - as an average of price relatives, 63-65
 - computation from frequency distribution, 313-16
 - confidence interval for, 221-24
 - as a decision-parameter, 62-66, 241-48
 - in enumerative studies, 66
 - estimator for, 199
 - expected value, 150
 - formula for, 66
 - of a population, 62-66
 - of a sample, 123-27
 - sampling distribution, 175-89
- Array, 67
- Assignable causes, 336
- Association
 - and causation, 406-8
 - degree of relationship, 403-6
 - dependent variable, 402
 - independent variable, 402
 - meaning of, 401-3
 - multiple relationship, 402
 - nature of relationship, 403-6
 - simple relationship, 402
- Attribute, 32-36
- Automatic recording devices
 - audimeter, 301
 - photoelectric cell, 301
 - program analyzer, 301
- Average; *see* Arithmetic mean
- Average deviation, 74-75

B

- Backman, Jules, 94
- Ballowe, James M., 343
- Bancroft, Gertrude, 288
- Bar chart, 37

- Bassie, V. Lewis, 486
 Beta risk (β -risk), 235
 Bias; *see* Procedural bias
 Biased estimator, 203-4
 Binary numbers, 325
 Binomial formula, 144-47
 Bivariate population, 35-36, 402, 435
 Blair, Morris Myers, 121
 Bratt, Elmer Clark, 486
 Brenner, H., 95
 Brinkman, J. S., 344
 Cross, Irwin D. J., 19
 Brunk, Max E., 393
 Built-in deficiency, 276-77
 Burns, Arthur F., 478-79
 Burr, Irving W., 275
- C**
- c*-chart, 351-54
 Call-backs, 285, 298-300
 Causation and association, 406-8
 Cause and effect, 406-8
 Census, 26, 88-90, 99-100, 276
 Centile, 51
 Chance; *see* Probability, Risk
 Chance causes, 336
 Chapin, F. Stuart, 288
 Chappell, Matthew N., 300
 Characteristic, 22-25
 Chebycheff, Pafnuti Lvovitch, 83
 Chunk, 103
 Cluster sampling, 118, 209-10
 Cochran, William G., 160, 400
 Coding, 305, 307-8, 314-16, 318-19
 Coefficient of correlation, 432-38
 Coefficient of determination, 435
 Collator, 322, 323
 Columbia Statistical Research Group, 264
 Column diagram, 39
 Common causes, 336
 Comparative experiments
 analysis of variance, 387-91
 defined, 370
 design, 382-84
 matched (paired) sample design, 382-86
 OC curve, 379
 several populations, 386-96
 two population means, 372-79
 two population proportions, 379-82
 Complete count, 12, 26; *see also* Census
 Complete information, 11, 21
 Completely randomized design, 393-94
 Conditional probability
 defined, 135
 dependent trials, 135
 rule, 135-37
 rule applied, 136-41
 Confidence coefficient, 221
 Confidence interval, 221-25, 427-28, 429
 Conklin, Maxwell R., 325
 Consequences, 6-10
 Consistency of data, 305, 457-58
 Consistent estimator, 199-201
 Constant seasonal, 477
 Consumer's risk, 258
 Continuing decision problem, 332-34
 Continuous variate, 32-33
 Control chart
 choice of, 350-51
 control limits, 337
 as a decision rule, 337
 frequency of sample selection, 359-60
 historical development, 333-34
 for mean ($\bar{x}$ -chart), 344-48
 number of defects (*c*-chart), 351-54
 for proportion defective (*p*-chart), 336-44
 p-chart versus $\bar{x}$ - and *R*-charts, 350-51
 R-chart, 344, 348-49
 rational subgroup, 358-60
 and risks of incorrect action
 probability limits, 355
 repeated sampling, 356-57
 3-sigma limits, 355
 types of, 355
 sample size, 360
 specification limits, 360-63
 3-sigma limits, 339, 355
 as a time series, 447-48
 Controllable causes, 336
 Convenience sampling, 102-3, 105-7
 Correlation
 bivariate normal distribution, 435
 coefficient of determination, 435
 critical values for sample coefficient of, 437
 estimator for coefficient of, 436-37
 model, 435
 nature of, 403-6
 population coefficient of, 432-35
 rank, 438-41
 sample coefficient of, 436-38
 Cowden, Dudley J., 121, 331, 368, 485
 Cox, Gertrude M., 400
 Cramér, Harald, 153
 Critical value, 238
 Croxton, Frederick E., 121, 331, 485
 Cyclical fluctuations, 454-55, 478-80
- D**
- d_2 , 346
 D_3 and D_4 , 346
 Dalleck, Winston C., 195
 Data
 external, 92-96
 internal, 91-92

- Data—*Cont.*
 - primary, 90–96
 - seasonally adjusted, 476–77
 - secondary, 90–96
 - Data accuracy
 - extreme values, 306
 - inconsistencies, 305
 - unlikely values, 305–6
 - unusual regularity, 306–7
 - Data collection
 - automatic recording devices, 301
 - direct observation, 297
 - mail questionnaire, 299–300
 - organization of, 302–5
 - panel technique, 300–301
 - personal interview, 298–99
 - telephone interview, 300
 - Data flow, 296–331
 - Data processing, 305–9
 - Data tabulation, 309–13
 - Decision making in the face of uncertainty, 2–6
 - Decision-parameter
 - defined, 51
 - measures used as, 53–83, 372, 379
 - sampling distribution of, 189–90
 - selection of, 83–84
 - states of the world, 51
 - statistic contrasted with, 125–26
 - Decision problems
 - estimation, 12–14, 194–231
 - nature of, 1–12
 - testing, 12–14, 232–74
 - Decision rule
 - in acceptance sampling, 259–61
 - alternate management objectives, 268–71
 - in comparative experiments, 374–75, 380–81
 - concept of, 15–16
 - in correlation problems, 437–38, 440
 - to discard data, 306
 - in estimation problems, 197–98*
 - operating characteristic curve, 248–57
 - in regression problems, 420–28
 - in testing problems, 236–48
 - Deming, W. Edwards, 48, 49, 208, 294, 331
 - De Moivre, Abraham, 78
 - Dependent variable, 402
 - Design of experiments
 - completely randomized, 382
 - efficiency of, 391
 - matched (paired) samples, 382–86
 - matched sample and completely randomized design contrasted, 386
 - nature of, 371–72
 - types of design
 - completely randomized, 393–94
 - Design of experiments—*Cont.*
 - types of design—*Cont.*
 - Latin Square, 395–96
 - randomized block, 394–95
 - Detlefsen, G. R., 95
 - Differences
 - of means, 372–73
 - of proportions, 379–80
 - of successive observations, 458–59
 - Discrete variate, 32
 - Dispersion; *see* Variability
 - Distribution, population, estimation of, 225–28
 - Distribution, sampling, 154–93
 - Double sampling, 118–19
 - Double sampling plan, 263–64, 266–68
 - Duncan, Acheson J., 153, 274, 368, 399
- E**
- Editing, 305
 - Edwards, G. D., 48
 - Efficient estimator, 203–5
 - Electronic computer, 323–26
 - Elementary unit, 22–25
 - Enumerative study, 28–31, 48
 - Equiprobable assumption, 130–31
 - Estimation
 - attained precision, 217–20
 - decision rule, 197–98
 - determining sample size
 - finite population, 214–15, 216
 - infinite population, 212–14, 216
 - estimator, 198–207
 - interval estimates, 197, 220–25
 - point estimate, 197
 - of population distribution, 225–28
 - problems, 12–14, 195–97, 210–17
 - uses of, 194–96
 - Estimator
 - for aggregate, 199
 - for arithmetic mean, 199
 - biased, 203
 - consistent, 199–201
 - defined, 198
 - efficiency of, 203–5
 - minimax, 205–7
 - for proportion, 199
 - ratio, 203–4
 - unbiased, 201–4
 - for variance, 201–2, 218–19
 - Expected value, 148–50, 161–62, 184–85
 - Exponential distribution, 81–82
 - Extreme values, 68–70, 306
- F**
- F distribution, use in analysis of variance, 391
 - F , tables of, 509
 - Factorials, 101

- Federer, Walter T., 393
 Field, 320
 Final report, 326-28
 Finite population, 29-30
 Finite population correction factor, 170-72, 187
 Fisher, Sir Ronald, 76, 371, 400
 Fisher, Willard C., 64
 Forecasting, 446-86
 Fraction defective, 339
 Frame, 25-28, 113-14
 Free-response question, 307
 Frequency distribution
 of attributes, 36-37
 bimodal, 44-45
 class intervals of, 310-12
 class limits of, 309-10
 classes of, 309, 311-12
 column diagram of, 39
 construction of, 309-12
 defined, 36
 histogram of, 39, 41-45
 J-shaped, 43-44
 multimodal, 44
 relative, 39-40
 skewed or asymmetrical, 42-44
 as a state of the world, 52-53
 symmetrical, 41-42
 unimodal, 44
 U-shaped, 44
 of variates, 37-40
 Frickey, Edwin, 478
- G**
- Galton, Sir Francis, 67
 Gauss, Karl Frederick, 78
 Gaussian distribution, 78; *see also* Normal distribution
 Gini, Corrado, 1
 Grant, Eugene L., 48, 274, 368
 Gray, Percy G., 290, 300
 Gretton, Owen C., 325
 Grouped data, 309
 Grubbs, F. E., 259
- H**
- Hansen, Morris H., 295
 Hirsch, Werner Z., 87, 153, 193, 230, 445
 Histogram, 39, 41-45
 Hochstim, Joseph R., 278
 Hooper, C. E., 300
 Hooperating, 300
 Huff, Darrell, 281, 282, 295
 Hurwitz, William N., 295
 Hypergeometric distribution, 170
 Hypothesis, 237-38
- I**
- Incomplete information, 10
 Independent trials, 141-47
 Independent variable, 402
Industrial Quality Control, 368
 Infinite population, 29-30
 Inspection, 257
 Interpreter, 322
 Interval estimation, 197, 220-25
 Interviewer
 bias, 288-89
 selection, training, and supervision, 303-4
- J**
- Judgment sampling, 102-4
- K**
- Kelly, Truman L., 86
 Kendall, M. G., 49, 112
 Key-punch machine, 322
 Kimball, A. W., 295
 Kish, Leslie, 291
 Kroll, Frank W., Jr., 366
- L**
- Lansing, John, 291
 Latin Square design, 395-96
LCL, 337
 Least-squares method, 417-18
 Legendre, Adrien, 417
 Lewis, E. E., 193
 Logarithmic scale, 451-52
 Logarithms, table of, 506-8
 Lot, definition of, 257
 Lot tolerance per cent defective (*LTPD*), 258
 Lower control limit (*LCL*), 337
LTPD, 258
- M**
- Machine tabulation, 319-26
 Mail questionnaire, 299
 Mandel, B. J., 87, 122, 485
 Mark sensing process, 322
 Mathematical probability; *see* Probability
 McCandless, Boyd, 285
 McCarthy, Phillip J., 19, 48, 445
 Mean; *see* Arithmetic mean
 Mean deviation, 74-75
 Measurement error, 286-87
 Medial average, 472, 474
 Median
 computation of, 67-68
 as a decision-parameter, 67-71
 defined, 67-68
 of a population, 67-71
 properties of, 69-70
 of a sample, 203
 Mid-value of a class, 312, 314-19
 Mills, Frederick C., 49, 331, 485

- Minimax estimator, 205-7
 Minimum inspection costs, 270
 Mitchell, Wesley C., 478-79
 Mode
 as a decision-parameter, 71
 of a population, 67, 71
 Moving average, 466-70
 Moving seasonal, 477
 Multiobservational problems, 31-32
 Multiple sampling plan, 263-64, 266-68
 Multiplier or calculator, 323
 Multivariate population, 36, 402
 Mutually exclusive outcomes, 132-33
 Myers, Robert J., 294
- N**
- National Bureau of Economic Research, 479-80
 Neiswanger, William A., 122, 331
 Neter, John, 49, 87, 122, 231, 274
 Newman, Joseph W., 106
 Neyman, J., 274
 No observations, 12, 88-90
 Nonparametric statistics, 83, 438-41
 Nonresponse, 282-86
 Nonsampling errors, 275-95
 Nonsense correlation, 406-8
 Normal curve, 165, 500-501; *see also*
 Normal distribution
 Normal deviates, 165, 502
 Normal distribution
 estimation of population model, 225-28
 as population model, 42, 77-83, 225-28
 as sampling distribution in comparative experiments, 373-74, 379-81, 384-85
 as sampling distribution of p , 160, 165-75
 as sampling distribution in regression problems, 419-27
 as sampling distribution of $\bar{x}$, 179-84
 and standard deviation, 77-83
 table of normal deviates, 502
 Normal equations, 417
 Null hypothesis, 238
- O**
- Observations
 biased, 286-91
 bivariate, 35-36
 of characteristics, 22-25
 continuous, 32-33
 discrete, 32
 multivariate, 36, 402
 numerical, 2
 qualitative, 32-33
 quantitative, 32-33
- Observations—*Cont.*
 in statistics, 10-12
 univariate, 35-36
 OC curve; *see* Operating characteristic curve
 One-tailed test, 242-43
 Open-ended question, 307-8
 Operating characteristic curve
 calculation of, 249-57
 in comparative experiments, 379
 defined, 248-49
- P**
- p -chart, 336-44
 Panel technique, 301
 Parameter, 52; *see also* Decision-parameter
 Parten, Mildred, 295, 307, 331
 Pearson, Karl, 67, 76
 Personal interview, 298-99
 Planning, the principle of, 14-16, 96
 Point estimate, 197
 Polya, G., 153
 Population
 of attributes, 33-35
 bivariate, 35-36, 402
 defined, 20-21
 existent, 49
 finite, 29
 finite twofold, 170-75
 hypothetical, 49
 infinite, 29
 multivariate, 36, 402
 univariate, 35-36
 of variates, 33-35
 Population model
 bimodal, 44-45
 bivariate normal distribution as, 435
 defined, 41
 J-shaped, 43-44
 multimodal, 44
 normal curve as, 42
 skewed or asymmetrical, 42-44
 symmetrical, 41-42
 U-shaped, 44
 unimodal, 44
 uses of, 45-46
 Precision
 accuracy contrasted with, 276-77
 attained, 217-20
 meaning of, 203
 Prestige bias, 290
 Primary data, 90-96
 Primary study, 90-96, 195-96
 Probability
 addition rule, 133
 binomial formula, 144-47
 conditional, 134-41

Probability—*Cont.*

- defined, 128–29
- distribution, 154
- estimation of, 129–31
- expected value and, 148–50
- independent trials, 135, 141–44
- mathematical, 127–34
- measure, 128, 132–34
- and measurement of risk, 140–41
- mutually exclusive outcomes, 132–33
- and the normal curve, 165–70
- theory, 122–53

Probability sampling

- advantages of, 103, 107
- defined, 103
- types of,
 - cluster sampling, 118
 - double sampling, 118–19
 - sequential, 119
 - simple random, 107–10
 - stratified, 116–18

Procedural bias

- accuracy vs. precision, 276–77
- control of, 276–77
- defined, 275
- sources of
 - arithmetical errors, 292
 - auspices of study, 289–90
 - built-in deficiency, 276–77
 - data flow, bias in, 291–92
 - improper formulation of problem, 277–79
 - interviewer bias, 288–89
 - measurement errors, 286–87
 - nonresponse, 282–86
 - observations, biased, 286–91
 - prestige bias, 290
 - questionnaire, 287–88
 - selection procedure, bias, 279–82

Producer's risk, 258

Proportion

- in analytical studies, 57
- confidence interval for, 225
- as a decision-parameter, 53–58, 236–41
- defined, 53–54
- in enumerative studies, 57
- estimator for, 199
- minimax estimator for, 205–6
- point estimate of, 199
- in a population, 50–51, 53–58
- in a sample, 125
- sampling distribution of, 156–64

Punch card, 319–21

Q

- Quality control, 333–34; *see also* Control chart

Questionnaire

- bias, 287–88
- described, 302
- fixed alternative questions, 307–8
- free-response or open-ended questions, 307–8

Quota sampling

- defined, 105
- stratified sampling contrasted with, 117–18

R

R-chart, 344, 348–49

Rand Corporation, The, 112, 503

Random digits, table of, 503–5

Random errors, 286–87

Random numbers, 111–16

Random process, 103, 107

Random sampling; *see* Probability sampling

Random variable

- defined, 124–25
- expected value of, 148–50
- statistic as a, 126–27

Randomization, 375

Randomized block design, 394–95

Range

- of a population, 74
- of a sample, 346–47

Range chart; *see* *R*-chart

Rank correlation, 438–40

Ratio estimate, 203–4

Ratio-to-moving-average method, 471–75

Recalls, 285, 298–300

Reciprocals, tables of, 488–99

Rectangular distribution, 81–82

Regression

- curvilinear, 405
- judgment applications of, 428–32
- linear, 404
- problems, nature of, 403–6
- sampling procedure for, 413–14
- scatter diagram, 405–6
- simple linear-regression model, 408–11

Regression line

- confidence interval estimates, 427–28, 429
- least-squares method, fitted by, 417–18
- for a population, 411
- slope of
 - defined, 411–12
 - estimator for, 414
 - standard error of, 419–20
- standard deviation of regression
 - defined, 413
 - estimator for, 418–19
- standard error of estimated *Y*-value
 - estimator for, 423–26
 - meaning of, 424

- Regression line—*Cont.*
 standard error of line
 estimator for, 422–23, 425–26
 meaning of, 422
 work sheet for computation of, 416
 Y-intercept
 defined, 411
 estimator for, 414
 Rejection number, 264–65
 Relative seasonal, 477
 Reproducer and summary punch, 323
 Rice, Stuart A., 289
 Risk
 in estimation problems, 194–231
 of incorrect decisions, 11–12
 measurement of, 140–41
 in testing problems, 232–74
 Roberts, Harry V., 19, 87, 153, 193, 231, 295, 400, 445
 Rosander, A. C., 193, 230
- S**
- Sample
 arithmetic mean, 124
 defined, 97
 as part of population, 23
 pattern of variability, 127
 proportion, 125
 selection, 88–122
 size, formulas for determining, 212, 214, 216, 241, 248
 standard deviation, 125
 variance, 125
 Samples, all possible, 100–102
 Sampling
 accuracy of, 98–99
 in analytical study, 29–30
 versus complete count, 99–100
 convenience, 102–7
 defined, 97
 economy of, 97, 98
 in enumerative study, 30
 infinite population, and, 99
 judgment, 102–7
 methods of
 cluster, 118, 209–10
 cost of, 207–10
 double, 118–19
 sequential, 119
 simple random, 107–10
 stratified, 116–18, 208–9
 versus no observations, 99
 probability, 12, 102–7
 quota, 105
 reasons for, 96–100
 risk, measurement of, 140–41
 timeliness of, 98
 Sampling distribution
 arithmetic mean
 development of, 175–84
 expected value, 184–85
 normal curve, 179–84
 sample size, 179–84
 sampling risk, 186–88
 standard error, 185–86
 decision-parameter, 189–90
 defined, 154–56
 difference in sample means, 373
 difference in sample proportions, 379–80
 hypergeometric, 170
 proportion
 binomial formula, 156–57
 expected value of p , 161–62
 finite population, 170–75
 infinite population, 156–64
 normal distribution, 160
 relation to population proportion, 157–58
 relation to sample size, 158–60
 sampling risk, 167–69, 172–74
 standard error, 162–64
 slope of regression line, 420–21
 Sampling error versus nonsampling error, 275–77; *see also* Procedural bias
 Savage, Leonard J., 19
 Scatter diagram, 405–6, 415–17
 Schedule, 298, 302
 Schweitzer, Henry, 94
 Seasonal movement
 absolute, 477
 adjustment for, 476–77
 constant, 477
 moving, 477
 ratio-to-moving-average method, 471–75
 relative, 477
 Secondary data
 defined, 90
 external, 92–96
 internal, 91–92
 principle of planning, 96
 reliability of, 96
 uses of, 91–96
 Secondary studies
 estimation problems, 196–97
 in general, 90–96
 operating characteristic curve, 256
 Sequential sampling, 119, 263, 265–68
 Shewhart, Walter A., 333
 Shubin, John A., 195
 Simmons, Willard A., 208
 Simple linear-regression model, 408–11
 Simple random sampling
 analytical studies, 110
 characteristics of, 108–9

- Simple random sampling—*Cont.*
 defined, 108
 enumerative studies, 108-9
 from finite population, 108-9
 from infinite population, 110
 Single sampling plan, 263-68
 Slope
 defined, 411-12
 estimator for, 414
 standard error of estimate of, 419-20
 Slutsky-Yule effect, 469-70
 Smith, B. Babington, 112
 Sorter and sorter counter, 322
 Specification limits, 361-63
 Sprowls, R. Clay, 19, 48, 274, 445, 486
 Spurious correlation, 406-8
 Square roots, tables of, 488-99
 Squares, tables of, 488-99
 Standard deviation
 computation from frequency distribution, 316-19
 as a decision-parameter, 75-77
 defined, 76-77
 and exponential curve, 81-82
 formula for, 76-77
 interpretation of, 77-83
 and the normal curve, 77-83
 of a population, 75-83
 and rectangular distribution, 81-82
 of a sample, 125
 Standard deviation of regression
 defined, 413
 estimate of, 418-19
 Standard error
 of difference in sample means, 373-74, 377-78
 of difference in sample proportions, 380-81
 of estimate; *see* Standard deviation of regression
 of estimated regression line, 422-23, 425-26
 of estimated Y-value, 423-25, 426
 of mean
 finite population, 185-86
 infinite population, 187
 and population size, 189
 population standard deviation, 189
 sample size, 189
 of proportion, 162-64
 finite population, 171-72
 infinite population, 163
 and π , 164
 and population size, 174-75
 and sample size, 163-64
 of slope of regression line, 419-20
 States of the world, 6-10; *see also* Decision-parameter
 Statistic
 decision-parameter contrasted, 125-26
 defined, 124
 as a random variable, 126-27
 Statistical control, 337-341
 and satisfactory product, 360-63
 Statistical decision rule; *see* Decision rule
 Statistical decisions by association, 401-45
 Statistical design of comparative experiments, 369-400
 Statistical experiment
 allocation of sample, 374-75
 defined, 369
 design of, 371-72
 multifactor, 392
 randomization, 375
 Statistical population; *see* Population
 Statistical table
 column heading, 312
 footnote, 313
 format, 310-11, 312-13
 source, 313
 stub, 313
 title, 312
 uses of, 309
 Statistics
 as a field, 2-3
 historical notes, 1, 62, 67, 76, 78, 83, 263, 333, 417
 as numerical data, 2-3
 as plural of statistic, 123-27
 Stock, J. Stevens, 290
 Stockton, John R., 121
 Stouffer, Samuel A., 67
 Stratified sampling
 defined, 116
 quota sampling contrasted, 117-18
 Suchman, Edward A., 285
 Systematic errors, 287
- T**
- t*, tables of, 509
 Table; *see* Statistical table
 Tabulator, 322
 Target population, 27-28
 Tchebycheff, Pafnutti Lvovitch, 83
 Tchebycheff's inequality, 82-83
 Teamwork, principle of, 16-17
 Telephone interview, 300
 Testing problem
 in acceptance sampling, 257-68
 alpha risk in, 235
 beta risk in, 235
 decision rule, elements of, 238
 defined, 12-14
 error of first kind, 235
 error of second kind, 235

- Testing problem—*Cont.*
 one-tailed, 242–43
 two-tailed, 242–43
 type I error, 235
 type II error, 235
 3-sigma limits, 339, 355
 Ties in ranks, 439
 Time series
 absolute changes in, 450–51
 cyclical fluctuations, 454–55, 478–80
 defined, 447
 and forecasting, 446–86
 graphic presentation of, 449–52
 moving average, 466–70
 random movements, 456
 relative changes in, 450–51
 seasonal movements, 455–56, 470–78
 serial correlations, 449
 Slutsky-Yule effect, 469–70
 trend element, 453, 458–66
 Tippett, L. H. C., 112, 399
 Tolerable error, 210–11
 Torray, Philip, 94
 Trend
 defined, 453
 fitting of, 460–64
 linear, 461–63
 nonlinear, 461–62
 projection of, 464–66
 statistical test for, 458–61
 Two-tailed test, 242–43
 Type I error, 235
 Type II error, 235
- U**
- UCL*, 337
 Unassignable causes, 336
 Unbiased estimator, 201–4
 Uncertainty
 in estimation problems, 194–231
 meaning of, 3, 10–12
 in testing problems, 232–74
 Uncontrollable causes, 336
 Unequal class intervals, 311
 United Nations Subcommittee on Statistical Sampling, 331
 Universe, 21; *see also* Population
 Upper control limit (*UCL*), 337
- V**
- Vance, Lawrence L., 274
 Variability
 as a criterion for decision, 72–73
 measures of
 average deviation, 74–75
 range, 74
 standard deviation, 75–83
 variance, 75–83
 in population, 52, 71–83
 in a process
 assignable-unassignable causes, 336
 controllable-uncontrollable causes, 336
 and decision making, 336–37
 special-common causes, 336
 Variance
 as a decision-parameter, 75–83
 defined, 76–77
 estimator for, 201–2, 218–19
 formula for, 76–77
 interpretation of, 77
 population, 75–83
 sample, 125
 Variate, 32–36
 Verifier, 322
- W**
- Wald, Abraham, 263
 Walker, Helen M., 67, 87
 Wallis, W. Allen, 19, 87, 153, 193, 231, 295, 307, 400, 445
 Wasserman, William, 49, 87, 122, 231
 Welch, Emmett H., 288
 Working population, 27–28
- X**
- $\bar{x}$ -chart, 344–48
- Y**
- Y*-intercept
 defined, 411
 estimator of, 414
 Yates, Frank, 48, 104, 331
 Yule, G. Udny, 49
- Z**
- z*, 165, 502

This book has been set on the Linotype in 11 point Modern No. 21, leaded 2 points and 10 point Modern No. 21, leaded 1 point. Chapter numbers are in 18 point Alternate Gothic caps and chapter titles in 24 point Alternate Gothic caps and lower case. The size of the type page is 27 by 46½ picas.
